ОТЧЁТ
РАБОЧЕЙ ГРУППЫ AAPOR
О НЕСЛУЧАЙНЫХ ВЫБОРКАХ

Фонд
Общественное
Мнение

ОТЧЁТ
РАБОЧЕЙ ГРУППЫ AAPOR
О НЕСЛУЧАЙНЫХ ВЫБОРКАХ

УДК 316.653
ББК 60.527

Отчёт рабочей группы AAPOR о неслучайных выборках: июнь 2013 / Американская ассоциация исследователей общественного мнения/ Пер. с англ. Д. Рогозина, А. Ипатовой. М. : Общероссийский общественный фонд «Общественное мнение», 2016. 170 с.

Уже более 60 лет для сбора данных и производства статистических выводов полстеры в основном применяют случайный отбор. Лишь недавно сложности с покрытием и неответами, ведущие к росту затрат на организацию опросов, подтолкнули профессиональное сообщество к размышлению об осмысленности привлечения методов построения неслучайных выборок. В отчёте AAPOR проанализированы слабые и сильные стороны различных методов неслучайного отбора и описаны условия, при которых различные исследовательские дизайны без применения случайной выборки могут быть полезны для распространения выводов (результатов) на более широкие совокупности. Книга рассчитана на широкий круг читателей, в том числе представителей некоммерческих организаций, СМИ, руководителей, социологов, других гуманитарных ученых, математиков, статистиков, демографов, аспирантов, студентов – на всех, кто проявляет интерес к вопросам социологических исследований и пытается понять природу общественного мнения.

ISBN 978-5-4465-1006-1

Оглавление

РАБОЧАЯ ГРУППА AAPOR

Рег Бейкер, сопредседатель Рабочей группы,
Market Strategies International
Дж. Майкл Брик, сопредседатель Рабочей группы, Westat
Ненси Бейтс, Бюро переписи США
Майк Батталья, Battaglia Consulting Group, LLC
Мик Купер, Мичиганский университет
Джил Девер, RTI International
Криста Гайл, Массачусетский университет в Амхерсте
Роджер Туранжо, Westat

БЛАГОДАРНОСТИ

Неоценимую помощь в работе над отчётом оказали наши коллеги, которые подготовили обзор и дали свои комментарии:

Роберт Борух, Уортонская школа бизнеса при Пенсильванском университете
Марио Каллегаро, Google
Митч Эггерс, Global Market Insite
Дэвид П. Фэн, Миннесотский университет
Линда Пиекарски, Survey Sampling International
Джордж Терханиан, Toluna
Ян Вернер, Jan Werner Data Processing
Клиффорд Янг, IPSOS

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Отчёты рабочих групп Американской ассоциации исследователей общественного мнения долгие годы задают стандарты качества опросной индустрии.

К настоящему времени (декабрь 2015 года) американские коллеги подготовили 14 отчётов по актуальным методологическим темам: онлайн-исследования, опт-ин-панели, методология электоральных прогнозов, защита персональной информации, опросы по мобильным телефонам, проблемы недостижимости и неответов и т. д. Каждому отчёту предшествовала огромная организационная и библиографическая работа, каждый представляет собой свод текущих достижений и перспектив развития, согласованных в группе ведущих специалистов по выбранной тематике.

Несмотря на огромную значимость методологической работы AAPOR для развития отрасли, на русский переведен лишь один отчёт Ассоциации – о больших данных (2015 года). Предлагаемый читателю отчёт о неслучайных выборках – второй и, очень надеемся, не последний шаг в направлении детального разбора мировых трендов, а значит, развития отечественной методологической культуры.

Трудно переоценить актуальность перевода отчёта о неслучайных выборках для России. Несмотря на то что в медиасреде представления результатов опросов сопровождаются упоминаниями о случайных ошибках выборки, а значит, де-факто говорится о доминировании случайных отборов в массовых опросах, подавляющее большинство отечественных исследований проводится на неслучайных выборках.

Этому способствуют три фактора.

Во-первых, за последние годы катастрофически снизился интерес людей к участию в опросах. Коэффициент неответов на уровне трети или четверти обращений считается не только приемлемым, но уже и целевым показателем для многих полстеров. Если в конце 1980-х годов на волне «перестройки, гласности и нового мышления» интервьюерам приходилось защищаться

от неотобранных в качестве респондентов людей – каждый желал высказать своё мнение, то теперь приходится уговаривать респондента участвовать в опросе и подстраиваться под его прихоти.

Во-вторых, за последние 20 лет сформировался заметный перекос в структуре себестоимости опросов. Затраты на полевые работы занимают куда меньшую долю, чем требуется для реализации полноценных случайных выборок. Оплачивая несколько сот рублей за полностью взятое интервью и никак не регулируя и не поощряя правильность выполнения интервьюерами отбора, исследователи подталкивают последних к переопределению разработанных правил на текущие, наиболее подходящие приёмы рекрутинга респондентов.

Наконец, в-третьих, отсутствует культура методологического сопровождения опросов, и как следствие, из опросной практики постепенно исчезают методические отчёты, описывающие механизмы сбора полевых данных. Если отсутствует запрос на оценку качества реализованной выборки, не регистрируются и не анализируются параданные, а вопросы надежности и валидности редуцируются до обобщённых, неверифицируемых суждений, полевые работы постепенно сводятся к наиболее простым, дешёвым и удобным для интервьюеров способам отбора респондента. Маршрутная выборка становится конформной, квотный отбор внутри домохозяйства переносится на квотирование в начале маршрута (то есть интервьюер ещё до обращения к членам отобранного домохозяйства начинает реализовывать квотное задание), происходят незначительные отклонения от правил отбора (фальсификации) или полевой материал подменяется полностью (фабрикации).

Поскольку большинство исследовательских компаний не проводят методический анализ качества полевых работ, поскольку транслируется лишь нормативный набор признаков систематической выборки, как правило, ограниченный несколькими страницами убористого текста, в среде профессионалов и публичном дискурсе произошла подмена понятий: закрыв глаза на реалии интервьюерской работы, исследователи стали именовать неслучайные выборки случайными. Это позволило, с одной стороны, в короткие сроки реализовывать сложные проекты с относительно небольшими

бюджетами, с другой – оставаться в научном дискурсе, претендуя на научно обоснованные выводы, построенные на материале, до последнего времени считавшемся в мире ненаучным.

Отчёт AAPOR о неслучайных выборках развенчивает миф об их ненаучности. Тем самым легитимируется российская традиция организации опросов, до этого находившаяся в тени не только публичных обсуждений, но и методической рефлексии. «Мы-то знаем, как обстоят дела на самом деле, но с текущей ситуацией ничего не поделать» – фраза, не раз звучавшая в кулуарных разговорах, теперь может быть вынесена за скобки опросной индустрии. Научность определяется не разработанными когда-то нормами и правилами отбора, не неукоснительным выполнением установленных регламентов и процедур, а развернутой системой регистрации происходящего и принципиальной открытостью методической информации. Неслучайно в большинстве документов, выпускаемых AAPOR (не исключение и настоящий отчёт), в качестве важнейшего требования к качеству данных выдвигается транспарентность методической информации, а два основополагающих документа, на которые должны ориентироваться все члены Ассоциации, – это Инициатива AAPOR по транспарентности данных (в том числе зафиксированная в Этическом кодексе ассоциации [The code of professional ethics and practices, 2015]) и Стандартные определения (правила регистрации и анализа методической информации) [Стандартные определения, 2005].

Работа над переводом отчёта о неслучайных выборках стала возможной благодаря пристальному вниманию к стандартам качества опросных данных, уделяемому в Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации. Мы признательны директору Института социального анализа и прогнозирования Т. Малевой за поддержку переводческой инициативы; коллективу коллег – Е. Гришиной, А. Бурдяк, А. Тындик, Е. Цацуре, Ю. Флоринской, Ю. Чумаковой, – своими критическими замечаниями, подчас сомнениями в надёжности данных социальных обследований стимулирующих нас к поиску новых аргументов, построению экспериментальных планов, наблюдению за реалиями сбора социальной информации и, как следствие, обращению к зарубежному опыту.

Организационную и финансовую поддержку, а также дружеское участие в столь сложном и непривычном в российском контексте методологическом труде нам уже долгие годы оказывает фонд «Общественное мнение». Мы признательны А. Ослону за неизменный интерес к методической стороне массовых опросов и неоценимую помощь в реализации настоящего перевода и иных методических авантюр, проводимых под эгидой возглавляемой им опросной компании. Сотрудники фонда – пожалуй, лучшие в России специалисты по выборочному исследованию А. Чуриков и Т. Османов – помогали нам не только определиться со сложными переводческими решениями, но и обогащали, восполняли пробелы в нашем профессиональном образовании. Без их участия мы никогда бы не решились на подобную работу.

В течение 2015 года, во второй половине которого перевод осуществлялся, мы регулярно встречались с коллегами по Лаборатории методологии федеративных исследований ИНСАП РАНХиГС, обсуждали текущие затруднения, анализировали выходящие в свет труды, намечали возможные варианты развития концептуального аппарата. Мы признательны Е. Вьюговской, Н. Галиевой, В. Картавцеву, К. Мануильской, Д. Сапонову за неоценимый вклад в развитие общей методологической рамки, в которой только и можно вести речь о стандартах качества.

Вместе с тем коллеги не несут ответственности за допущенные в переводе ошибки и неточности, которых может быть немало. Первые шаги по осознанию текущей опросной ситуации, отказу от нормативного мышления в пользу действительно научного наблюдения за реалиями жизни опроса зачастую осуществляются вслепую. Не исключение и переводческие решения. Здесь возможен и выбор не вполне точной терминологии, и семантические просчёты, и ошибки, связанные с приписыванием иного контекста отрывистым и понятным зарубежным коллегам примерам. Ошибаться нормально. И нормально исправлять ошибки. Поэтому мы призываем коллег, которые прочтут этот отчёт, обсуждать недочёты и делиться своим видением ситуации. Мы ждём от вас отзывов, комментариев, критических замечаний по следующим адресам: nizgor@gmail.com (Дмитрий Рогозин) и ipatova_anna@mail.ru (Анна Ипатова).

Итак, приступайте к чтению отчёта. И пусть подчас оно будет трудным и неувлекательным, мы надеемся, что этот документ станет ещё одним кирпичиком в нашем общем здании методологического осмысления опросной технологии.

Литература:

1. Отчёт AAPOR о больших данных: проектная группа AAPOR, 12 февраля 2015 / Л. Джапек, Ф. Крейтер, М. Берг и др.; Американская ассоциация исследователей общественного мнения / Пер. с англ. Д. Рогозина, А. Ипатовой, Е. Вьюговской. М., 2015.

2. Стандартные определения: систематическое описание диспозиционных кодов и коэффициентов результативности для массовых опросов. 3-е изд. 2004 / Американская ассоциация исследователей общественного мнения / Пер. с англ. Д. М. Рогозина, Е. М. Киселева // Социологический журнал. 2005. № 2. С. 78–119.

3. The code of professional ethics and practices / American Association for Public Opinion Research; Revised version 2015. 30 November.

Дмитрий Рогозин
декабрь 2015

АННОТАЦИЯ

Исследователи, регулярно проводящие массовые опросы, используют разные методы сбора данных и производства статистических выводов. Уже более 60 лет для этих целей в основном применяют случайный отбор. Лишь недавно сложности с покрытием и неответами, ведущие к росту затрат на организацию опросов, подтолкнули профессиональное сообщество к размышлению об осмысленности привлечения методов построения неслучайных выборок – как минимум при определённых условиях.

Известно огромное количество неслучайных дизайнов, которые включают исследования сходных случаев (исследования «случай-контроль», case-control studies), клинические испытания, оценочные исследования, перехватывающие исследования (intercept surveys), опт-ин-панели (панели согласившихся принимать участие в онлайн-опросах) и многие другие. И хотя эти дизайны повсеместно используются в других областях знания, исследователями общественного мнения они до сих пор подробно не изучены.

Осенью 2011 года Исполнительный совет Американской ассоциации исследователей общественного мнения (AAPOR Executive Council) поставил задачу «изучить условия, при которых различные исследовательские дизайны без применения случайной выборки могут быть полезны для распространения выводов (результатов) на более широкие совокупности». Главная особенность статистического вывода заключается в том, что он требует некоторого теоретического основания и явного набора допущений для вынесения оценок и суждений об их точности. Мы считаем, что методы сбора данных и формулирование оценок без теоретических оснований недопустимы при производстве статистических выводов.

В отчёте мы анализируем слабые и сильные стороны разных методов неслучайного отбора, приводя теоретические и в некоторых случаях практические доказательства. Мы не претендуем на исчерпывающее исследование всех методов или доскональный обзор литературы даже об одном из них. Однако надеемся, что нам удалось обозначить наиболее распространённые методы и рассмотреть их в сбалансированном и объективном ключе.

ОБЗОР ОТЧЁТА

Раздел 1 отчёта – краткое введение в тему, раздел 2 – описание опыта, который за последние годы приобрели исследователи, обращаясь к случайным и неслучайным методам построения выборки; цель этого экскурса – представить эволюцию идей, которые лежат в основе выборочных исследований и являются предметом текущего интереса AAPOR.

В разделе 3 мы обращаемся к некоторым базовым проблемам формирования неслучайных выборок, акцентируя внимание на сложности производства статистических выводов. Мы также описываем некоторые методы, не относящиеся непосредственно к задачам, поставленным в отчёте, поскольку для них отсутствуют теоретические основания или компоненты дизайна выборки.

В разделах 4, 5 и 6 подробно описываются принципы неслучайных методов, которые исследователи могут принимать во внимание. В каждом из этих методов – свой подход к организации выборки. Один подход – выравнивание (сопряжение) выборки (sample matching) – долгие годы применялся в неэкспериментальных исследованиях[1] и теперь активно используется в опт-ин-панелях. Второй подход – сетевой отбор (network sampling) – включает выборку, управляемую респондентами (respondent driven sampling, RDS); он активно используется для отбора редких или труднодостижимых групп, когда применение методов случайного отбора невозможно. В последнем из этих трёх разделов обсуждается набор post hoc корректировок, которые предлагаются как способы снижения смещений в оценках для неслучайных выборок. Эти корректировки обращаются к дополнительным данным, которые используют для исправления смещений в ходе отбора, а также смещений других типов. Корректировка по степени склонности (propensity score adjustment – PSA) – пожалуй, наиболее известная из этих техник.

[1] Экспериментальным планам традиционно противопоставляются исследования, построенные на наблюдениях (observational studies). Их особенность состоит в том, что измерения производятся в естественной среде. Исследователь пытается контролировать значимые переменные, не вмешиваясь во взаимодействие социальных объектов. Поэтому такой тип исследований часто называют неэкспериментальными, квазиэкспериментальными, натурными и т. д. – *Прим. перев.*

В разделах 7 и 8 обсуждаются методы определения точности оценок и концепт соответствия для применения (fitness for use). Случайные выборки имеют отчётливые критерии качества, сфокусированные вокруг концепта общей ошибки исследования (total survey error, TSE). Неслучайные выборки не могут быть полноценно представлены в этой теоретической рамке, потому мы рассматриваем некоторые возможные альтернативы общей ошибке исследования. Вероятно, это и есть самая насущная потребность, если неслучайные методы начнут активно внедряться в практику массовых опросов. Концепт соответствия для применения также рассмотрен подробно, и, по всей видимости, он может быть весьма релевантным как для неслучайных, так и для случайных выборок. Однако эта область требует дальнейших интенсивных исследований.

В разделе 9 представлены выводы нашего отчёта. Кратко их опишем.

- **В отличие от ситуации со случайными выборками, не существует какой-либо общей теоретической рамки, которая адекватно представляла бы все способы неслучайных отборов.** Пока это коллекция методов, и чрезвычайно сложно, если вообще возможно, описать общие свойства всех подходов к построению неслучайных выборок.

- **Может оказаться плодотворным представление о различных подходах к построению неслучайных выборок как некотором континууме ожидаемой точности оценок.** Исследования, расположенные на нижней и верхней границах континуума, легко опознать по усилиям, затраченным на контроль выборки, и post hoc корректировкам. Сложность в размещении методов между двумя экстремумами и в оценке рисков, сопровождающих статистические выводы, сделанные по результатам таких опросов. Риск зависит от знания предмета и применяемых техник.

- **Транспарентность (прозрачность) – неотъемлемая методическая сущность.** При применении методов неслучайного отбора, в отличие от ситуации со случайными выборками, есть большие

ограничения в описании методов составления выборки, сбора данных и производства статистических выводов. Большинство онлайн-исследований не сопровождаются информацией, адекватно описывающей их методологию.

- **Производство статистических выводов и для случайных, и для неслучайных исследований основано на некоторых допущениях о применимости используемых моделей расчётов.** Эти допущения должны быть чётко описаны, с максимальной полнотой должны быть описаны и эффекты, которые возникают в результате этих допущений и могут влиять на точность оценок.

- **Наиболее перспективные неслучайные методы основаны на моделях, в которых проблемы, связанные со статистическим выводом, преодолеваются как на этапе отбора, так и на этапе оценки.** В подходах, основанных на моделировании (model-based approaches), как правило, предполагается, что ответы обобщаются в соответствии со статистической моделью (например, все наблюдения имеют одни и те же среднее и дисперсию). В модели обычно стремятся использовать важнейшие вспомогательные переменные, чтобы сделать её более удобной и пригодной для применения. Как только модель сформулирована, для получения статистических выводов об оцениваемых параметрах совокупности применяются стандартные статистические процедуры оценивания, такие как вероятностные, или байесовские техники.

- **Одна из причин нечастого использования в массовых опросах методов, основанных на моделировании, заключается в том, что разработка релевантных моделей и тестирование заложенных в них допущений весьма сложны и затратны по времени, к тому же требуют хорошей статистической квалификации.** Допущения должны быть проверены для всех ключевых оценок – модель, хорошо работающая с одними оценками,

может давать сбои с другими. Достижение простоты, присущей случайным выборочным методам в производстве множественных оценок, весьма проблематично для методов неслучайного отбора.

- **Соответствие цели – важнейший концепт для оценки качества исследования, но его применение в опросном дизайне требует дальнейших разработок.** Организации, которые работают со случайными выборками, пытаются поддерживать баланс между такими качествами исследования, как релевантность, точность, своевременность, доступность, интерпретируемость и согласованность. Аналогичные усилия требуются и для неслучайных выборок.

- **Методы построения выборки, используемые в опт-ин-панелях, в последнее время бурно развиваются, поэтому исследователь, которому необходимо оценить валидность получаемых данных, должен скорее обращать внимание на методы отбора, чем на саму панель.** Пользователи опт-ин-панелей могут применять различные методы отбора, процедуры сбора данных и техники оценивания. Прежние подходы к организации неслучайных выборок из панелей имеют невысокую релевантность в сравнении с современными подходами.

- **Если неслучайные выборки станут широко использоваться в практике массовых опросов, потребуется более согласованная теоретическая рамка и сопутствующий набор измерителей для оценки их качества.** Одно из ключевых достоинств случайного отбора – это комплект готовых измерителей и конструктов (таких, как общая ошибка исследования), которые определяют понимание качества и позволяют выявлять источники ошибок. Применение этого набора для оценки неслучайных выборок неэффективно из-за иных оснований отбора. Общепризнана острая необходимость в исследованиях, нацеленных на развитие показателей качества, в том числе оценки смещений и точности полученных на неслучайных выборках значений.

- **Неслучайные выборки хорошо себя зарекомендовали в электоральных опросах, но нет столь же определённых свидетельств их качества в других сферах, в том числе в комплексных обследованиях с целью изучения различных феноменов.** Исследования, нацеленные на получение ограниченного числа оценок по заданному набору результатов, требуют контроля за малым набором сопутствующих переменных. Однако зачастую исследования направлены на широкий круг предметных областей и, соответственно, требуется большое количество оценок, подразумевающих большой набор сопутствующих переменных.

- **Неслучайные выборки могут быть пригодны для производства статистических выводов, но валидность этих выводов зависит от уместности допущений, заложенных в модель, и от того, насколько отклонения от допущений влияют на конкретные оценки.** На протяжении всего отчёта мы проводим мысль, что для любого метода построения неслучайных выборок необходимо развивать теоретическую основу, опирающуюся на эмпирическую оценку метода. В этой оценке должны проверяться пригодность допущений в различных случаях и для разных статистических расчётов. В отчёте подчеркивается, что выравнивание выборки – это один из методов, уже имеющих сконструированную для оценочных исследований теоретическую основу, которая должна быть модифицирована и адаптирована под требования массовых опросов. Некоторые исследователи приступили к решению этой задачи. Методы постопросных корректировок, применяемые в неслучайных выборках, являются плодом усилий, предпринятых ранее исследователями случайных выборок. Хотя в некоторых случаях они могут быть пригодны и эффективны, требуется дополнительное рассмотрение механизмов смещения при отборе. Мы считаем, что повестка по развитию метода должна включать эти требования.

1 Введение

При организации массовых опросов исследователи используют разные методы сбора данных и производства статистических выводов. Уже более 60 лет большинство из них придерживаются парадигмы случайного отбора. Лишь недавно сложности с покрытием и неответами, которые приводят и к росту затрат на организацию опросов, дали повод для размышлений об осмысленности привлечения методов построения неслучайных выборок – как минимум при определённых условиях.

Дизайн неслучайных выборок разнообразен – это исследования с контролем кейсов (исследование сходных случаев, case-control studies), клинические испытания, оценочные исследования, перехватывающие исследования (intercept surveys), опт-ин-панели (панели согласившихся принимать участие в онлайн-опросах) и многие другие. И хотя эти дизайны широко используются в других областях знания, исследователями общественного мнения они до сих пор подробно не изучены. Ввиду того что неслучайная выборка имеет ограниченное применение, основные допущения, необходимые для того чтобы построить надёжные выводы в опросах такого рода, также до конца не изучены.

Осенью 2011 года Исполнительный совет Американской ассоциации исследователей общественного мнения (AAPOR Executive Council) поставил задачу «изучить условия, при которых различные исследовательские дизайны без применения случайных выборок могут быть полезны для распространения полученных оценок на более широкую совокупность». Мы понимаем, что термин «статистический вывод» имеет много значений и толкований. В данном отчёте мы определяем его как набор процедур, необходимых для оценки характеристик изучаемой совокупности и предоставляющих способ

измерения надёжности этих оценок. Главная особенность статистического вывода заключается в том, что он требует некоторого теоретического основания и явно выраженного набора допущений для производства оценок и вынесения суждений об их точности. Мы считаем, что методы сбора данных и формулирование оценок без теоретических оснований недопустимы при производстве статистических выводов.

Возможно, некоторые читатели ждут, что в данном отчёте будет идти речь о сравнении методов случайной и неслучайной выборок, об их сильных и слабых сторонах. Эти читатели, скорее всего, будут разочарованы. Мы обошли стороной методы построения случайных выборок в неидеальных условиях, равно как и сравнение случайной и неслучайной выборок. Безусловно, важно, насколько случайная выборка действительно является случайной в условиях низкого покрытия или высокой доли неответов, но рабочая группа не ставила перед собой цели найти ответ на этот вопрос.

В наши намерения входило изучение сильных и слабых сторон разных методов построения неслучайных выборок, рассматривая теоретические и в определённой степени эмпирические основания. Мы не претендуем на исчерпывающее исследование всех возможных методов или доскональный обзор литературы даже об одном из них. Однако надеемся, что нам удалось как минимум обозначить наиболее распространённые методы и рассмотреть их в сбалансированной и объективной манере.

Использование неслучайных выборок широко распространилось с ростом распространения онлайн-опросов. Как правило, основной источник для выборки в онлайн-исследованиях – группа людей, рекрутированных заранее и согласившихся принимать участие в опросах. В данном докладе мы используем термин «опт-ин-панель» (панель согласившихся принимать участие в онлайн-опросах) для обозначения онлайн-панели из людей, которые были рекрутированы без применения вероятностного отбора. Способы построения таких панелей различны. За последние годы исследователи, работающие с опт-ин-панелями, стали пользоваться техниками, выходящими за пределы простого квотирования. Важно, что формирование панели основывается не на одном методе отбора, а на бесчисленном

и разнообразном множестве этих методов. При оценке результатов таких исследований необходимо в большей мере сосредоточить внимание на выборочных методах формирования собственно панели.

Мы понимаем, что для многих сотрудников AAPOR термины «научное исследование» и «случайная выборка» являются практически синонимами, то же самое можно сказать о многих членах нашей рабочей группы. И если вы непреклонны в убеждении, что статистический вывод невозможен без случайного отбора, или если вы твердо уверены, что метод построения выборки не имеет значения для распространения результатов исследования на всю совокупность, данный доклад вряд ли будет иметь для вас большую ценность. Так или иначе, в докладе мы попытались быть непредвзятыми. В этой связи уместно вспомнить следующие слова Киша: *«Большие шаги вперёд в наиболее успешных науках – астрономии, физике, химии – были сделаны и делаются до сих пор без вероятностного отбора. Построение статистических выводов в такого рода исследованиях основывается на субъективном суждении о наличии адекватного, автоматического и естественного случайного распределения в совокупности… Не существует понятного правила для решения вопроса, когда необходим вероятностный отбор и какую цену за него надо заплатить. Вероятностный отбор для обеспечения случайности является не догмой, а стратегией, особенно когда речь идет о больших множествах»* [Kish, 1965, p. 28–29].

Цель данного доклада не в том, чтобы установить «чёткое правило». Правильнее говорить о первом шаге, который, как мы надеемся, вызовет широкую дискуссию в профессиональном сообществе. Мы считаем, что такая дискуссия остро необходима, и эта необходимость назрела уже давно, поскольку мы столкнулись с проблемой адаптирования наших методов к уже сильно изменившейся и продолжающей меняться реальности. Такая задача возникает в долгой истории исследований общественного мнения не впервые и возникнет ещё не раз. И как всегда это и происходило, решение этой задачи сделает нашу профессию сильнее.

2 История вопроса

Случайная выборка в течение многих десятилетий доминировала в исследованиях, но ни в коем случае не была единственной и не всегда преобладала. Не будем подробно описывать всю историю вопроса, тем более хорошие обзоры такого рода уже существуют (например, [Frankel & Frankel, 1987; Brick, 2011]), а только кратко затронем основные её этапы и приведем наиболее распространённые аргументы в нынешних спорах о преимуществах дизайнов исследований, основанных на случайном и неслучайном отборах.

Основная задача выборочных исследований заключается в том, чтобы делать надёжные и точные выводы относительно широких групп населения. Это также называют «репрезентативностью», хотя Киш [Kish, 1995] писал, что данный термин недостаточно точен. В целом отправной точкой для выборочных исследований считается труд Неймана[2]. До этого при проведении опросов использовались два основных подхода (см., например, [Kiaer, 1895-6; Yates, 1946]). Первый заключался в выборе репрезентативного поселения (исходя из некоторых допущений) и далее в проведении переписи в этом поселении. Второй представлял собой целевой отбор, то есть в выборку включались территории и единицы отбора на основании заданных критериев или параметров по типу квотной выборки. Комментарий Неймана к работе 1929 года, описывающей использование целевой выборки в Италии

[2] Ежи Нейман (Neyman, Jerzy) – польский и американский математик, статистик, член Национальной академии наук США. Работал в Калифорнийском университете. Он первым ввел в статистические исследования понятие доверительного интервала. При этом Киш отмечает важность работ двух советских математиков: А.А. Чупрова, чьё имя запечатлено в названии размещения Чупрова–Неймана (1923), и А.Г. Ковалевского, чью работу «Основы теории выборочных методов» (1924) Киш считал первой книгой о выборках [Kish, 2003, p. 6]. – *Прим. ред.*

[Gini, Galvani, 1929], актуален и в современных спорах о неслучайной выборке: «Сравнение выборки с населением всей страны показало, что хотя средние значения в использованных семи контрольных переменных расходятся незначительно, средние значения других переменных, которые не контролировались, часто не совпадают. Соответствие других статистик, таких как частотные распределения и т. п., происходит ещё реже» [Neyman, 1934, p. 585]. Цитата из труда Неймана не только демонстрирует слабость таких практик того времени, как целевая выборка и повальная перепись, но и показывает, какие идеи стали основой теории и практики научных выборочных исследований. По мнению Киша, Нейман установил «триумф случайной выборки из многих малых и неравных кластеров, стратифицированных для лучшей репрезентативности» [Kish, 1995, p. 9].

В то время как «триумф случайной выборки» наблюдается при сборе официальной статистики национальными статистическими службами, в опросах общественного мнения распространения она не получает до тех пор, пока на выборах 1936 и 1948 годов не происходят два грандиозных публичных провала опросов, основанных на неслучайной выборке. В 1936 году журнал Literary Digest разослал 10 млн предвыборных почтовых бюллетеней (straw poll ballots), на основании 2,3 млн заполненных предсказал победителя выборов – и ошибся (см. [Bryson, 1976; Squire, 1988]). Прогнозы же на основании проведённого Джорджем Гэллапом предвыборного опроса с использованием квотной выборки оказались верны.

На выборах в 1948 году все трое лидирующих полстеров того времени – Кроссли, Гэллап и Роупер – использовали метод квотной выборки, и все они неправильно предсказали победителя.

Мостеллер и его коллеги [Mosteller et al., 1949] провели анализ результатов предвыборных исследований 1948 года, оказавший значительное влияние на развитие опросной технологии. Они выявили множество источников ошибок в опросах. В качестве одного из потенциальных источников ошибок было использование квотной, а не случайной выборки. В докладе по результатам анализа отмечалось: использование квотной выборки привело к тому, что интервьюеры иногда выбирали более образованных и состоятельных

людей, и это повлекло за собой смещение не в пользу Трумэна. Хотя в докладе саму квотную выборку не «обвиняют» в неправильных прогнозах, её применение было поставлено под сомнение – главным образом потому, что она не позволяет оценить надёжность результатов опроса избирателей. Последствия не заставили себя ждать – вскоре Гэллап начал применять методы, основанные на случайном отборе, и его примеру последовали остальные полстеры. В общем, опросная индустрия приняла допущение, что небольшой хорошо спроектированный выборочный опрос может дать более точные результаты, чем гораздо больший по объёму, но менее проработанный.

Тем не менее прогнозы, сделанные на основании неслучайных методов отбора, иногда бывают точны. Часто забывают, что Literary Digest проводил опросы перед каждыми выборами с 1920 по 1936 год, и всегда правильно предсказывал победителя. Однако один публичный провал привёл к закрытию журнала в 1938 году (или во всяком случае этому поспособствовал).

В обзоре истории выборочных исследований в США Франкель и Франкель [Frankel and Frankel, 1987, p. 129] писали: «После 1948 года споры сторонников квотной и случайной выборки закончились, и случайная выборка стала для США предпочтительным методом». Тем не менее ещё несколько десятилетий квотную выборку продолжали использовать параллельно со случайными методами отбора, особенно в маркетинговых исследованиях, где её применяют и по сей день, и даже в университетской среде. Например, в первые годы при американском Общем социальном обследовании (GSS, General Social Survey[3]) использовалась именно квотная выборка, а переход на случайный отбор был совершён в 1977 году. Комбинирование случайного отбора (для первичных и вторичных единиц отбора) и квотной выборки (для домохозяйств или людей в такого рода единицах) было распространено в Европе многие десятилетия и до сих пор используется в некоторых странах (см., например, [Vehovar, 1999]). Замена не принявших участия

[3] Американское Общее социальное обследование (General Social Survey, GSS) включает сбор демографических сведений и данных об отношении к тем или иным явлениям постоянных жителей США. Впервые опрос был проведён по телефону Национальным центром изучения общественного мнения Чикагского университета (National Opinion Research Center at the University of Chicago) в 1972 году и далее проходил практически ежегодно. С 1994 года он проводится раз в два года. Интервью в среднем занимает около 45 минут. – *Прим. перев.*

в опросе респондентов – также всё ещё относительно распространённая практика за пределами Северной Америки (например, [Vehovar, 1995]). Схожим образом, понятие репрезентативной выборки как выбора поселения или района, подробного его изучения и дальнейшего распространения полученных выводов на более широкие группы населения до недавних пор существовало в России и других странах.

Исторически так сложилось, что основные аргументы против случайных выборок заключались в их стоимости и временных затратах. Этот аргумент был довольно убедительным, когда самым распространённым методом опроса были личные интервью. После введения в 1970-х годах телефонных опросов со случайным набором номера (RDD, random digit dial) (см. [Glasser & Metzger, 1972]) баланс методов изменился. Распространение методов случайного отбора на телефонные опросы помогло решить проблему покрытия выборок, основанных на телефонных справочниках (directory samples), а также сократить значительные усилия, которые были необходимы для построения таких выборок и проведения опроса. При помощи случайного набора номера опрос может быть сделан относительно недорого и быстро, с приемлемым покрытием всего населения и — по крайней мере так было в первое время — с относительно низким уровнем неответов.

Появление дозвона со случайным набором номера (RDD) привело к широкому распространению выборочных методов в области политических опросов, маркетинговых исследований и в академической среде. Этот бум продолжался вплоть до резкого увеличения доли домохозяйств, где есть только мобильный телефон [Lavrakas et al., 2007], приведшего к проблемам с ошибкой покрытия. В то же время многолетнее падение уровня ответов, отчасти связанное с повсеместным использованием телефонов для проведения массовых опросов и в особенности для телемаркетинга, поставило вопрос об ошибке неответов (см., например, [Brick, Williams, 2013; Curtin, Presser, Singer, 2005]).

Такие факторы, как быстро растущая стоимость опросов с использованием традиционных методов, основанных на случайном отборе (личные и телефонные опросы), снижающийся уровень ответов и растущие опасения

относительно покрытия в телефонных опросах, привели к тому, что большие надежды стали возлагаться на онлайн-исследования, особенно когда сильно выросла интернет-аудитория [Couper, 2000]. Однако невозможность развивать методы отбора респондентов для онлайн-исследований, аналогичные случайному набору номера, породила альтернативные подходы, основанные на неслучайных методах отбора, прежде всего опт-ин-панели. Такие панели дают возможность получать опросные данные от большого числа респондентов в краткие сроки и по относительно низкой стоимости. Имея доступ к миллионам потенциальных респондентов, подгруппам или людям, которые обозначили в своём опросном профиле[4] определённые характеристики, можно отбирать участников специализированных социальных обследований. Первые доводы за использование такого рода панелей были основаны на их размере (напоминает доводы журнала Literary Digest), более высоком по сравнению с телефонными опросами уровне ответов (по крайней мере так было в первое время) и на возможности собирать вспомогательные данные (auxiliary data) для корректировки. Популярность таких опросов, причём не только в маркетинговых исследованиях, но и в политических опросах и даже в научных исследованиях, привела к необходимости более пристально изучить надёжность их результатов (эти проблемы подробно рассмотрены в [Baker et al., 2010]). Тем не менее, как уже было отмечено рабочей группой Американской ассоциации исследователей общественного мнения по вопросам онлайн-панелей (AAPOR Task Force on Online Panels), такого рода опросы несомненно представляют ценность для некоторых видов исследований, но исследователям «следует избегать опт-ин-панелей с неслучайным методом отбора респондентов, когда основная задача заключается в точной оценке характеристик всей генеральной совокупности <…> при использовании таких источников для выборки следует избегать претензии на „репрезентативность“».

Итак, перед социальным исследователем в очередной раз встала сложная и интересная задача. Как отмечали Франкель и Франкель, «до 1960 года

[4] Имеется в виду profile surveys – чтобы войти в базу подписчиков, потенциальный респондент (получатель рассылки) должен ответить на несколько вопросов, то есть пройти небольшой пробный опрос. Данные, полученные в ходе этого опроса, составят профиль респондента на сайте. – *Прим. перев.*

отказы не рассматривались как угроза построению выводов на основе выборочных исследований. Главная проблема заключалась в отсутствии респондента дома, что можно было решить посредством повторных контактов» [Frankel, Frankel, 1987, p. S133]. За последние 10 лет рост числа попыток контакта (например, большее число звонков) стал не только недостаточным для решения проблем неответа, но и финансово нерациональным. Проблема низкого уровня ответов в выборках со случайным отбором и обусловленные этим смещения (ошибки неответов) служат основным аргументом за использование альтернативных подходов.

Мы не хотим сказать, что такие альтернативы, как роботизированные телефонные опросы (robo-calls) или онлайн опт-ин-опросы не имеют своих проблем неответов или покрытия. Но сторонники этих подходов утверждают, что привычная практика случайного отбора, где для минимизации ошибок в выводах необходимо иметь хорошее покрытие всех слоёв населения и высокий уровень ответов, сегодня настолько сложна и материально затратна в реализации, что этот метод подходит только для очень хорошо финансируемых и социально важных исследований, например таких, которые проводят национальные службы статистики. Они также утверждают, что если при соотнесении респондентов (которых получилось опросить) с изучаемой совокупностью использовать подходящий метод и правильные параметры, можно получить валидные выводы. Этот спор обычно характеризуют как спор между подходами, один из которых основан на проектировании (design-based approach), а другой – на моделировании (model-based approach).

Гроувз в своей книге также задался вопросом о неответах, дав одной из глав провокационное название: «Почему, имея высокий уровень неответов, мы используем случайный отбор?» [Groves, 2006]. Он отметил, что исследования с неслучайным отбором нагружают аналитиков двойной работой: необходимостью корректировать полученные результаты как с учётом неслучайного отбора, так и с учётом неответов. Для таких конструкций,

как аксесс-панель (access panels[5]), где возможность участия в панели ограничивается принадлежностью к определённой группе населения, вопрос о покрытии тоже актуален. Как недавно заметил Брик, «неответы, недостаточное покрытие населения и ошибки измерения являются примерами тех практических вопросов, которые мешают чистоте выводов в исследованиях со случайным отбором» [Brick, 2011].

Основной посыл нашей работы такой: само по себе использование в опросе случайной выборки не делает его результаты валидными и надёжно отражающими мнения тех групп населения, которые он должен представлять. Как и использование при проведении опроса произвольных методов не означает автоматически, что его результаты не заслуживают внимания или являются ошибочными.

Существует ряд подтверждённых случаев, когда при помощи неслучайного отбора были получены результаты, которые оказались не менее, а то и более точными, чем результаты опросов, проведённых при помощи случайного отбора. Особенно важно, что это происходило в области предвыборных опросов [Abate 1998, Snell et al., 1998; Taylor et al., 2001; Harris Interactive, 2004, 2008; Twyman 2008; Vavreck, Rivers 2008; Silver, 2012]. Аналогичным образом существует мнение, что альтернативные методы, как, например, выравнивание неслучайной выборки, могут быть столь же точны, сколь и методы случайного отбора. Такое происходит при выравнивании выборки (например, [Rivers, 2007]), при корректировке по степени склонности (propensity score adjustment – PSA), когда для этого используются подходящие переменные (например, [Terhanian and Bremer, 2012]) или же когда выполнены допущения для выборки, управляемой респондентами (respondent driven

[5] Аксесс-панель, по сути, является своего рода базой потенциальных респондентов, то есть содержит список тех людей, которые на постоянной основе участвуют в опросах за вознаграждение. Их можно назвать и «профессиональными респондентами». Это не панель в чистом виде, так как респонденты всё время участвуют в разных исследованиях, где их выбирают на основании указанных характеристик. Понятия «аксесс-панель» и «опт-ин-панель» – довольно близки, однако в настоящем отчёте второй термин используется для обозначения таких панелей, участники которых были рекрутированы без применения случайного отбора, тогда как понятие аксесс-панели шире и включает онлайн-панели всех видов. – *Прим. перев.*

sampling (например, [Heckathorn, 1997]). В теории, если допущения полностью выполнены (как это происходит при случайном отборе), полученные оценки должны быть несмещёнными.

Подведём итоги. Споры о возможности распространять выводы опросов, проведённых при помощи неслучайных методов отбора, на всю совокупность, а также противопоставление, с одной стороны, цены и времени, с другой – качества не новы. Технологический прогресс (в особенности использование интернета) привел к развитию новых методов (как предсказывали Франкель и Франкель в 1997 году), и число опросов с их использованием возросло, что только подогревает споры. Тем не менее сама эта задача в разных формах стоит перед нами с самых первых дней проведения выборочных опросов.

Основной вопрос заключается в уровне доверия и величине доверительного интервала, в который укладываются результаты опроса. Подход, основанный на проектировании выборки, при одновременном использовании моделей для корректировки неполного покрытия и неответов дает некоторую гарантию от смещения выборки. Подход с неслучайными методами отбора в большей степени полагается на применимость модели и в большинстве случаев на выбор, доступность и качество переменных, использованных для отбора респондента и корректировки полученных результатов.

Конечно же, опросы – это не только способ оценки параметров совокупности и, в отличие от случайной выборки, не существует единой концептуальной рамки, которая объединила бы все формы неслучайной выборки. В следующем разделе мы рассмотрим использование различных методов неслучайного отбора в разных условиях. Так или иначе, речь идет о выборочных исследованиях (или их альтернативах), где напрямую ставится цель распространить описания и аналитические выводы на генеральную совокупность, соответственно, наиболее уместной в данном контексте выглядит дискуссия о построении статистических выводов на основании случайных и неслучайных методов отбора.

3 Введение в методы построения неслучайных выборок

В предыдущем разделе говорилось, что в сравнении со случайной выборкой, где есть единая концептуальная рамка для проведения отбора респондентов и построения статистических выводов обо всей изучаемой совокупности, для всех форм неслучайного отбора такой рамки нет. Вернее, существует широкий круг методов построения выборки, из которых одни технически сложны и неоднозначны, а другие просты и прямолинейны.

В данном докладе мы исходим из посыла, что метод построения выборки надёжен только в том случае, если статистические выводы из выборки можно экстраполировать на большую изучаемую совокупность. Метод должен включать набор процедур для проведения оценки характеристик изучаемой совокупности, а также предлагать некоторую систему измерения надёжности этих оценок. (Например, при случайной выборке мы можем оценить среднее или суммарное значение для всей целевой совокупности и установить доверительные интервалы, которые показывают надёжность этих оценок). Таким образом, принципиально важно, чтобы метод имел под собой некую теоретическую основу или набор допущений, позволяющих производить оценку и определять точность этих оценок. Конечно, любая концептуальная рамка для построения выводов, включая случайную выборку, имеет допущения, которые не всегда выдерживаются на практике; и величина отклонений от этих допущений является критической при оценке

применимости этих статистических выводов. Мы считаем, что методы сбора данных и расчёта показателей, не имеющие под собой теоретической основы, *не подходят* для получения статистических выводов. Конформная[6] выборка является одним из таких методов. В докладе мы не будем говорить о конформной выборке именно по этой причине – из-за отсутствия теории, но для полноты картины ниже коротко её опишем.

3.1. КОНФОРМНАЯ ВЫБОРКА

Во многих научных дисциплинах социального толка конформная выборка является основным методом отбора респондентов. Например, хотя психологи иногда используют данные, полученные посредством национальных репрезентативных случайных выборок, намного чаще их исследования основаны на конформной выборке из студентов колледжа. В середине 1980-х годов Дэвид Сирс [David Sears, 1986] выразил беспокойство по этому поводу. Он проанализировал статьи, опубликованные в трёх ведущих журналах по социальной психологии того времени, и обнаружил, что «в 1980 году 75 % статей этих журналов основаны только на данных, полученных от студентов <…> а большинство авторов статей (53 %) утверждают, что рекрутировали студентов, посещающих курс психологии». Когда он исследовал статьи в тех же журналах пять лет спустя, картина не изменилась. И хотя построение выводов, основанных на нерепрезентативных выборках, может быть ошибочным, как и утверждал Сирс, психологи продолжили использовать конформные выборки в большинстве своих исследований. Сирса больше волновала совокупность, из которой выбирались участники психологических экспериментов, нежели метод отбора, но очевидно, что даже сама совокупность студентов старших курсов колледжа, скорее всего, представлена в психологических экспериментах нерепрезентативно. Участники психологических экспериментов уже прошли самоотбор различными способами: от решения идти в конкретный колледж, посещать определённые классы до желания пройти отбор и принять участие в конкретном эксперименте.

[6] Мы выбрали перевод термина convenience sample как «конформная выборка» по соображениям, изложенным в статье: *Д. М. Рогозин*. Конформная выборка в торговых центрах // Социологический журнал. 2008. № 1. С. 22–48. – *Прим. перев.*

Использование конформных выборок далеко не ограничивается областью психологии (см., например, [Presser, 1984]). Некоторые виды судебных исследований (litigation research) основаны на перехватывающих выборках в торговых центрах (mall-intercept samples) [Diamond, 2000]; они популярны в случаях нарушения прав торговых марок, где изучаемая совокупность состоит из потенциальных покупателей продукта. Во многих таких случаях торговые центры – удобные места для поиска представителей изучаемой совокупности. В рандомизированных экспериментальных планах (randomized trials) в области экономики и образования испытуемые также часто отбираются неслучайным образом.

Значительное число медицинских исследований основано на неслучайной конформной выборке – зачастую это пациенты, к которым у исследователей есть непосредственный доступ. Купер [Couper, 2007] перечислил ряд медицинских исследований (объект которых варьируется от социального тревожного расстройства до язвенного колита), где для оценки состояния здоровья населения использовались онлайн-выборки; почти во всех случаях это были выборки добровольцев. Наконец, многие исследователи общественного мнения, использующие случайную выборку, также прибегают к конформной выборке в случае с фокус-группами или когнитивным тестированием анкеты. Таким образом, использование конформной выборки широко распространено даже среди исследователей, признающих превосходство случайных методов отбора в других условиях.

Определение конформной выборки. Большинство учебников по построению выборки не содержат формального определения конформной выборки, включая этот метод в число других неслучайных методов отбора. Потому начнём с определения: конформная выборка (convenience samples) – это вид неслучайной выборки, где в качестве первоочерёдного фактора выступает простота поиска или рекрутирования потенциального респондента. По названию понятно, что выбор респондентов основан скорее на удобстве (для исследователя), чем на каком-либо формальном дизайне выборки. Некоторые распространённые виды конформной выборки – это опросы в торговых центрах, выборка добровольцев (volunteer samples, стихийная выборка), потоковая выборка (river samples), выборка для некоторых

эмпирических исследований, а также некоторые выборки методом снежного кома. Далее мы кратко остановимся на каждом из типов конформной выборки, однако необходимо отметить, что на практике эти методы иногда могут приводить к формированию выборок, которые не подходят под определение конформных.

Перехватывающие опросы в торговых центрах (mall intercepts). При проведении таких опросов интервьюеры пытаются привлечь к участию покупателей или других прохожих в одном или нескольких торговых центрах. Как правило, ни торговый центр, ни респонденты не проходят случайный отбор, хотя некоторые систематические методы могут применяться для определения, к кому обращаться в конкретном торговом центре. Например, интервьюеры могут подходить к каждому N-му человеку, проходящему мимо конкретного места в торговом центре. И всё же в большинстве опросов в торговых центрах упор делается на быстрый набор необходимого числа респондентов за небольшую стоимость (с некоторой видимостью объективности). В результате выбор торговых центров и респондентов зачастую происходит бессистемно, а допущения, необходимые для распространения результатов на более широкую совокупность, оказываются не в приоритете. В большинстве судебных исследований (litigation studies) принадлежность выбранных для участия респондентов к изучаемой совокупности является принципиально важной, и чтобы установить, соответствует ли человек определённым критериям (например, для выявления потенциальных покупателей какого-то определённого типа продукта), могут применяться детализированные скрининговые вопросы. И хотя перехватывающая выборка опросов в торговых центрах может не позволить исследователям распространять количественные оценки и делать выводы обо всей совокупности, которую эта выборка должна представлять, как правило, суды считают такие опросы намного более ценными и информативными, чем их альтернативу – простой выбор нескольких людей, которые говорят о своих реакциях на указанный «раздражитель».

Панели добровольцев (panels of volunteers). Как уже было сказано, выборки добровольцев (volunteer samples, стихийные выборки) широко распространены в социальных науках, медицинских и маркетинговых исследованиях.

Как правило, люди соглашаются принять участие в каком-то конкретном исследовании, но иногда их включают в панель на некоторое время, а затем просят принять участие и в других исследованиях. Потребительские панели в маркетинговых исследованиях практикуются как минимум 50 лет, чаще всего в виде опросов по почте [Sudman and Wansink, 2002]. В последние годы рекрутируется большое число опт-ин-онлайн-панелей (opt-in web panels) для заполнения онлайн-анкет, каждый месяц участники панелей получают приглашения для прохождения большого числа онлайн-опросов [Couper, Bosnjak, 2010]. Такие опт-ин-панели были предметом исследования в более раннем отчёте рабочей группы Американских исследователей общественного мнения (AAPOR task force report) [Baker et al., 2010], где сделан вывод, что в целом результаты, полученные посредством такой выборки, менее надёжны, чем результаты, полученные с применением случайной выборки.

С течением времени исследователи, которые доверяли результатам опт-ин-панелей, осознали недостатки такого источника построения выборки. Сегодня растёт число работ, посвящённых методам корректировки возможных смещений в этих панелях, делающим результаты более точными и полезными. Эти методы будут подробнее рассмотрены в разделах 4 и 6.

Потоковые выборки (river samples). Потоковая выборка – это подход к построению выборки, основанный на использовании интернета, где респонденты отбираются на один опрос или включаются в панель повторяющихся опросов в течение долгого времени [GfK Knowledge Networks, 2008; Olivier, 2011]. В потоковой выборке в качестве потенциальных респондентов чаще всего выступают люди, посещающие сайты, где вывешено приглашение принять участие в опросе. Для того чтобы заинтересовать посетителей сайтов и вызвать у них желание пройти опрос (или даже стать участником опт-ин-панели), используются такие средства привлечения внимания, как всплывающие окна, гиперссылки и баннеры.

Отбор респондентов в потоковой выборке имеет два аспекта. Первый: необходимо решить, какие сайты будут выступать в качестве кластеров. Эти кластеры отличаются от тех, которые используются при случайной выборке, где представители генеральной совокупности относятся к какому-либо одному

кластеру. Здесь кластер – это интернет-сайт, посетители которого рекрутируются. Потоковая выборка, основанная на одном интернет-сайте, с большой долей вероятности будет состоять из индивидов, схожих (то есть однородных) по различным демографическим переменным, установкам и прочим характеристикам. Этого следует избегать, если требуется репрезентация широких групп населения, и на практике для формирования потоковой выборки обычно используется много сайтов. Для опроса всего взрослого населения может использоваться довольно большое число сайтов, и это снижает необходимость учитывать природу посещений каждого из них, при этом выборка получается разнородной (heterogeneous sample). В опросе, цель которого – изучить определённую группу людей, внешняя информация может помочь принять решение, какие именно сайты использовать. В теории, для улучшения покрытия среди изучаемой совокупности можно построить стратифицированную выборку интернет-сайтов. На практике сайты обычно отбираются из соображений компромисса между стоимостью и ожидаемым числом потенциальных респондентов, хотя иногда для отбора используется таргетирование по демографическим признакам (demographic targeting).

Второй аспект: чтобы определить, подходят ли для опроса люди, которые захотели в нём участвовать, может потребоваться скрининг. При отборе участников для конкретного исследования можно использовать самые разные скрининговые (отборочные) параметры. Если стоит задача включить человека в панель, его могут попросить заполнить более подробную анкету для определения профиля участника панели.

Таким образом, потоковая выборка – это основанная на согласии интернет-пользователей техника отбора (opt-in web-based sampling technique), которая исторически является видом конформной выборки. В последние годы с целью улучшения репрезентативности выборок особый упор делается на использование более формальных методов отбора.

Исследования, основанные на наблюдениях (observational studies). Исследования, основанные на наблюдениях, ставят своей целью проверку гипотез (обычно причинных гипотез) о медицинских или социальных феноменах без проведения рандомизированных контролируемых

экспериментов (controlled randomized experiments). (Более подробно исследования, основанные на наблюдениях, рассмотрены в следующем разделе). Во многих таких исследованиях тоже используется выборка добровольцев, хотя иногда они основаны на случайных методах отбора.

Рассмотрим известное Фремингемское исследование сердца[7] (Framingham Heart Study) – основанное на «репрезентативной выборке» исследование заболеваний сердечно-сосудистой системы (ЗССС), а также роли холестерина, курения и физической нагрузки в их развитии. Город Фремингем (штат Массачусетс) был выбран главным образом из соображений удобства (он был приблизительно нужного размера, имел всего одну больницу и актуальный справочник адресов жителей). В панель были включены как добровольцы, так и жители, отобранные при помощи систематического метода из городской переписи. Таким образом, дизайн выборки представлял собой некий гибрид, где место опроса было выбрано из соображений удобства, часть респондентов отобрана случайным методом, а другая часть состояла из «самовыбранных» добровольцев, которых добавили для того, чтобы увеличить общий размер выборки (подробнее см. первое описание исследования, на тот момент находящегося только на стадии разработки: [Dawber, Meadors, Moore, 1951]). Во многих клинических испытаниях (экспериментах) используют схожую комбинацию случайного и неслучайного отбора, выбирая больницы или врачебные практики неслучайно, а отдельных пациентов – уже с использованием случайных и неслучайных методов.

Выборка методом снежного кома (snowball sampling). Последний вид конформной выборки, о котором стоит сказать, – это выборка методом снежного кома. Этот метод появился как метод для отбора внутри сетей (sampling networks) и изначально не являлся видом конформной выборки. Коулман [Coleman, 1958] первым использовал эту технику как метод для изучения социального окружения человека. Например, человека могут попросить назвать своего лучшего друга, у которого потом возьмут интервью и также

[7] Фремингемское исследование сердца (Framingham Heart Study) – одно из наиболее известных и продолжительных медицинских исследований. Оно было начато в 1948 году по инициативе Общественной службы здоровья США и на сегодня проводится уже на третьем поколении. См. подробнее http://www.framinghamheartstudy.org/. – *Прим. перев.*

попросят назвать его или её лучшего друга. Гудман [Goodman, 1961] показал, что строгая версия этого метода с использованием случайного отбора, которой он дал название «выборка методом снежного кома», обладает необходимыми статистическими свойствами. Тем не менее в последующем этот метод обычно применялся в виде снежного кома с неслучайным отбором для поиска представителей труднодоступных или «скрытых» групп людей. Во многих таких более поздних исследованиях стартовые точки (the seeds) выборки методом снежного кома определялись в изучаемой популяции конформным образом, а не случайно (что ставил своей целью Гудман). Однако позже Хекаторн [Heckathorn, 1997] ввел особый вариант выборки методом снежного кома, который называется «выборка, управляемая респондентами» (respondent-driven sampling, RDS). Такой подход делает возможным построение специализированных выборок (specialized sampling), а также предлагает надёжные допущения, которые позволяют получить практически несмещённые оценки. Выборка, управляемая респондентами, наравне с другими видами выборки методом снежного кома или сетевых выборок будет рассмотрена более подробно в разделе 5.

3.2. УГРОЗЫ ДЛЯ СТАТИСТИЧЕСКИХ ВЫВОДОВ

Как мы уже отмечали, конформная выборка – только один из многочисленных видов неслучайных выборок, и её основным отличием является удобство построения. Как и все неслучайные методы отбора, конформная выборка имеет несколько потенциальных источников смещений, но в отличие от других неслучайных методов её приверженцы в большинстве случаев не предпринимают серьёзных попыток их исправить. Тем не менее эти смещения показывают, какие неточности в целом характерны для построения статистических выводов из неслучайных выборок.

Рассмотрим всё обследуемое методом опроса население, которое называют изучаемой совокупностью. В Фремингемском исследовании сердца, например, исследователи намеревались описать всё взрослое население в возрасте от 30 до 50 лет (на 1 января 1950 года) или, как минимум всё американское

взрослое население в этой возрастной группе. Соответственно, изучаемая совокупность, по-видимому, не ограничивалась жителями Фремингема. Существует значительная разница между изучаемой совокупностью, с одной стороны, и фактически отобранными для исследования людьми – с другой. Эта проблема – общая для большинства неслучайных выборок: группа, из которой осуществлялся отбор, вероятно, является малой и нерепрезентативной частью совокупности, которая интересует исследователя. (При использовании случайного отбора могут возникать аналогичные проблемы, когда основа выборки не предполагает полного покрытия целевой совокупности, так что всегда существует некоторый риск ошибки покрытия, даже в высококачественной случайной выборке. Но в этом случае доля населения, которое не представлено, в разы меньше, чем в конформной выборке.) Для неслучайных выборок лучше называть эту проблему «ошибкой невключения» (exclusian bias), чем «ошибкой покрытия» (coverage bias), поскольку подавляющая часть изучаемого населения, скорее всего, не имеет возможности попасть в выборку.

Отметим ещё один аспект: неслучайные выборки чаще всего состоят из добровольцев, и эти добровольцы могут не очень хорошо репрезентировать даже то население, из которого фактически осуществлялся отбор, не говоря уже о более широкой совокупности. При использовании случайной выборки вероятность отбора определяется исследователями и может быть включена в процесс оценки (посредством взвешивания). Тогда как в случае с выборками добровольцев вероятность участия определяется самими добровольцами и фактически неизвестна. Вероятное смещение при отборе (selection bias), отражающее систематические расхождения между добровольцами и не-добровольцами по значимым переменным, может быть существенным [Bethlehem, 2010].

Последняя помеха для статистических выводов из неслучайных выборок заключается в том, что коэффициент участия (зависящий от числа приглашённых для участия в исследовании) обычно очень низок. Исследования, основанные на опт-ин-панелях, не позволяют определить правильный коэффициент ответов (response rate); вместо этого приходится показывать *коэффициент участия* (participation rates), то есть долю тех, кто в конечном

итоге заполнил анкету, от числа участников панели, получивших приглашение на опрос. Коэффициент участия в онлайн-опросах, основанных на опт-ин-панелях, часто выражается одноразрядным числом (менее 10 %).

Несмотря на то что три перечисленные проблемы необязательно распространяются на все неслучайные выборки, во многих случаях они имеют место. Представим себе выборку участников опт-ин-панели, которых попросили пройти специализированный онлайн-опрос. Участники панели могли быть рекрутированы с относительно малого числа сайтов, что фактически исключает большинство представителей любой целевой совокупности, которая интересна исследователям, в частности тех, у кого нет доступа в интернет. Только малая часть тех, кто получил приглашение присоединиться к панели, могли согласиться, и только малая часть тех «панелистов», которых пригласили пройти специализированный опрос, могли принять в нём участие. Таким образом, окончательный набор ответов может иметь огромные смещения, связанные с невключением в совокупность, отбором и неучастием в опросе. И как мы отметили ранее, в выборке добровольцев сложно оценить итоговую вероятность отбора, необходимую для корректировки, и, соответственно, сложно включить её в процедуру взвешивания. Проводимые апостериорные корректировки основаны на сравнении полученных характеристик с ожидаемыми, а вовсе не на вероятности отбора.

3.3. ОЦЕНИВАНИЕ

При помощи неслучайных выборок иногда пытаются оценить характеристики совокупности, но чаще их используют для других целей. Например, подавляющее большинство психологических экспериментов не ставят своей задачей оценку средних показателей или пропорций какой-то конкретной совокупности, что характерно для опросов, основанных на случайной выборке. Вместо этого их цель – определить, действительно ли различия между двумя (или более) экспериментальными группами отличны от нуля. В других случаях не строится никаких выводов, основанных на количественном анализе (например, при формировании выборки для участия в фокус-группе).

Вне зависимости от предполагаемого использования данных необходимы некоторые допущения, позволяющие строить статистические оценки и определять их дисперсию. Слишком часто исследователи, использующие неслучайные методы отбора, строят количественные выводы, обращаясь с данными так, как будто они были получены при помощи простой случайной выборки. Такой подход существенно упрощает анализ данных, подсчёт стандартной ошибки и оценку уровня значимости (significance tests). Он предполагает, что метод отбора респондентов можно не учитывать. Надёжность допущения, что метод построения выборки не важен на стадии анализа полученных данных, – это предмет серьёзного спора.

Процедуры корректировки. Некоторые исследователи предпочитают не игнорировать механизмы построения выборки и получения ответов, а компенсировать потенциальные смещения, связанные с невключением, отбором и неучастием, используя какую-либо из процедур взвешивания. Например, чтобы привести данные, собранные опт-ин-панелью, в соответствие с известными распределениями населения, могут применяться веса, что является хоть какой-то попыткой учесть недопредставленность одних групп населения и перебор других. Распространённый метод такой коррекции, постстратификация (post-stratification) [Kalton, Flores-Cervantes, 2003], более подробно рассматривается в разделе 6. При постстратификации для выборки подбираются такие веса, чтобы ее параметры после взвешивания совпадали с соответствующими параметрами изучаемой совокупности в каждой ячейке, сформированной перекрёстной матрицей (cross-classifying) двух или более категориальных вспомогательных переменных (categorical auxiliary variables). Например, веса могут приводить параметры выборки в соответствие со значениями совокупности для каждой комбинации признаков «пол», «регион» и «возрастная категория». В случайных выборках начальный вес отдельного элемента совокупности (респондента) (case's initial weigh) рассчитывается обратно пропорционально вероятности его отбора, а потом умножается на поправочный коэффициент (adjustment factor). Для каждой ячейки перекрестной матрицы проводится отдельное выравнивание (хотя ячейки могут объединяться, чтобы не допустить чрезмерных

корректировок[8]). В случае с опт-ин-панелями начальные веса иногда просто приравниваются к единице. Общие параметры совокупности, необходимые для корректировки выборки, часто определяются только приблизительно на основании большого опроса, например Обследование американского общества (American Community Survey, ACS[9]) или Текущего опроса населения (Current Population Survey, CPS[10]). При помощи постстратификации можно избежать смещений, связанных с проблемами невключения или покрытия, если в каждой ячейке, которая подвергается перевзвешиванию, вероятность участия респондентов в опросе не влияет на искомые переменные, другими словами, если респонденты и нереспонденты в каждой конкретной ячейке имеют одинаковые распределения по анализируемым переменным. Это иногда называют «несистематическими случайными пропусками» (missing at random, MAR) [Little, Rubin, 2002].

Посредством метода выравнивания выборки (sample matching) делается попытка получить такую онлайн-выборку (web sample), которая изначально соответствует набору ключевых параметров изучаемой совокупности [Rivers, Bailey, 2009], а не выравнивается по ним постфактум. Например, когда из участников панели создается подвыборка для рассылки приглашений к участию в опросе, она формируется таким образом, чтобы ее структура совпадала со структурой изучаемой совокупности по некоторому набору вспомогательных переменных (auxiliary variable). В качестве этих вспомогательных переменных могут выступать стандартные демографические характеристики (скажем, половозрастные категории по регионам), а также установки респондентов, такие как политические взгляды или восприятие

[8] Ячейки перекрёстной матрицы объединяются для того, чтобы избежать экстремальных поправочных коэффициентов – слишком больших или, наоборот, близких к нулю. – *Прим. ред.*

[9] Обследование американского общества (American Community Survey, ACS) – один из наиболее масштабных проектов Бюро переписи США (за год опрашивается более 3 млн респондентов), который проводится на постоянной основе. Опрос проводится главным образом по почте и сосредоточен на занятости, уровне образования, доходе, миграции, происхождении, инвалидизации и характеристиках домохозяйства. Собранные данные широко используются в государственном и частном секторах.

[10] Текущий опрос населения (Current Population Survey, CPS) проводится Бюро переписи США совместно со Статистическим управлением Министерства труда США (Bureau of Labor Statistics, BSL). В рамках исследования ежемесячно собирается статистика о числе трудоустроенных и безработных граждан США. Впервые опрос был проведён в 1940-е годы, и на сегодня ежемесячно в нём принимают участие порядка 60 000 домохозяйств.

инноваций. Оставшиеся различия между структурой (make-up) выборки и структурой совокупности могут быть потом скорректированы при помощи перевзвешивания (подробнее см. раздел 6) — по крайней мере в некоторой степени. В теории, выравнивание выборки (sample matching) может повлиять на смещения так же, как и постстратификация, поскольку при выравнивании выборка изначально строится в соответствии с набором характеристик совокупности, а при постстратификации выборка приводится в соответствие с ними постфактум. Тем не менее неответы могут привести к тому, что получившаяся после опроса выборка не будет соответствовать известным параметрам населения, и хотя изначально она была выравнена, ей ещё потребуется дополнительная корректировка после опроса – постстратификация (о чем пишут Риверс и Бейли [Rivers, Bailey, 2009]). Методы выравнивания выборки более детально описаны в разделе 4.

Эффективность процедур корректировки (effectiveness of adjustment procedures). Несмотря на то что в принципе данные методы корректировки могут работать, снижая некоторые или все смещения оценок, полученных посредством неслучайной выборки, основной вопрос заключается в том, насколько они эффективны на практике. По крайней мере в восьми работах поднимался этот вопрос (см. обзор [Tourangeau, Conrad, Couper, 2013]).

Во всех исследованиях, направленных на оценку эффективности процедур корректировки, использовались похожие методы. Исследователи начинали с результатов онлайн-опроса или имитировали результаты онлайн-опроса путём выделения из телефонного или личного опроса подмножества респондентов, имеющих доступ в интернет. Затем они сравнивали оценки из реальных или сымитированных онлайн-опросов с некоторым набором контрольных значений. Контрольные значения могли браться из калибровочного исследования (calibration study) (личного или телефонного опроса, который шёл параллельно по случайной выборке), или же из всей выборки, если респонденты онлайн-опроса были подмножеством респондентов, имеющих доступ в интернет. В качестве контрольных значений также могли использоваться оценки из «внешних» опросов (таких как Текущий опрос населения США или любой другой источник, заслуживающий доверия; подробнее см. [Yeager et al., 2011]).

Несмотря на различия в способах получения контрольных значений и в специфике изучаемых корректировочных стратегий, в целом исследования показывают, что корректировка полезна, но решает проблему только частично. Неизвестно, смогут ли перевзвешивания или другие процедуры позволить аналитикам делать точные оценки для всего населения, основываясь на неслучайных выборках, таких как опт-ин-панели. Тем не менее исследования продолжаются, и особое внимание в них уделяется методам, позволяющим определить более широкий (по сравнению с описанным) спектр переменных, по которым можно производить корректировку. В отдельных случаях, о которых мы расскажем в разделе 7, исследователи могут смириться со смещениями, при условии что они не слишком велики для поставленных целей.

Оценки, основанные на данных экспериментов (estimation based on data from experiments). Как мы уже отмечали, многие проводимые психологами исследования основаны на конформной выборке, как и многие методологические эксперименты. Для экспериментов основной вопрос состоит в том, отличаются ли две или более группы друг от друга по одной или нескольким выходным переменным. Обычно это проверяется статистическим тестированием – значимо ли отличается от нуля смоделированный параметр (например, разность между средними значениями переменных в группах). Статистическое обоснование для тестирования значимости базируется на предположении, что участники были случайно распределены по экспериментальным группам. Уже говорилось, что в такой ситуации исследователь может поддаться искушению и невольно распространить сколько-нибудь значимые статистические различия на широкую совокупность. Однако данные, полученные в результате эксперимента, не дают основания для такого рода обобщений, если, конечно, до случайного распределения по экспериментальным группам не проводился случайный отбор из более широких групп населения. Такое происходит крайне редко. В результате исследователи, как правило, больше увлечены сопоставимостью экспериментальных групп, которые они могут контролировать, нежели тем, насколько хорошо эти группы репрезентируют большую совокупность. В классической терминологии, введённой Кэмпбеллом и Стэнли [Campbell, Stanley, 1963], исследователи, проводящие эксперименты, уделяют больше внимания внутренней валидности эксперимента, а не его внешней валидности или репрезентативности.

Является ли обоснованным такое пренебрежение условиями отбора? Другими словами, надёжно ли предположение, что результаты воздействия на участников эксперимента окажутся такими же и для генеральной совокупности? Здесь возможны варианты. Результаты эксперимента, скорее всего, будут внешне валидными, если соблюдено одно из двух условий. Во-первых, может случиться, что смещения, связанные с использованием конформной выборки, малы. Психологи иногда утверждают, что для базисных психологических процессов, затрагивающих память или восприятие, оценки экспериментов несильно зависят от смещения, связанного с отбором. Люди есть люди, и с учётом некоторых процессов между ними больше сходства, чем различий. Однако последние исследования показали, что этот аргумент используется слишком часто и существует много доказательств различий в тех характеристиках, которые считаются универсальными [Henrich, Heine, Norenzayan, 2010]. Во-вторых, смещения могут быть большими, но так или иначе уравновешиваться и сводить различия к нулю. Опять же, обоснованность этого предположения часто неоднозначна.

Тем не менее большая разница в смещениях между экспериментальными группами является общей проблемой, и это приводит к переоценке значимости для генеральной совокупности того эффекта, который был выявлен в ходе эксперимента. Например, эксперимент, в котором будет применён, скажем, альтернативный дизайн анкеты в онлайн-опросе, может иметь сильное воздействие на онлайн опт-ин-панель, но при этом в гораздо меньшей степени воздействовать на генеральную совокупность, члены которой имеют меньший опыт общения с онлайн-опросами, чем большинство участников панели, и потому они менее чувствительны к особенностям дизайна. Хуже того, может случиться, что результаты в экспериментальной и контрольной группах смещены относительно средних значений в совокупности не только по величине, но и в противоположных направлениях. В таком случае получится, что оценки, выведенные на основании эксперимента, не просто завышены или занижены, но и ошибочны: вместо ожидаемого увеличения значений контролируемых параметров в ходе экспериментального воздействия на изучаемую совокупность будет наблюдаться их уменьшение, и наоборот, вместо уменьшения – увеличение. Не до конца ясно, как часто такая ситуация возникает на практике.

3.4. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ ВЫВОДОВ БЕЗ ВЫБОРОЧНОГО ИССЛЕДОВАНИЯ

За последние годы появилось несколько интересных и инновационных подходов, которые используют самостоятельно возникающие и уже доступные данные для измерения характеристик населения или прогнозирования будущего поведения. В целом они отличаются от других методов, описанных в данном отчёте, двумя важными вещами. Во-первых, часто им вообще не требуется отбор респондентов, поскольку они оперируют большим накопленным объёмом данных, и имплицитно подразумевается, что большой объём снижает любые потенциальные смещения. Во-вторых, чтобы избежать оплаты сбора данных, в них практически не используются анкетирование или прямой опрос респондентов об их отношениях и поведении – эти сведения получаются из других источников.

Такие техники можно распределить по трём широким категориям: исследование социальных медиа (social media research), крауд-технологии (wisdom of crowds) и большие данные (big data). Каждую из этих групп мы обсудим ниже, хотя и без серьёзной оценки, поскольку, исходя из наших определений, они выпадают из любой концептуальной рамки выборочного исследования.

Исследование социальных медиа (social media research). Эта группа методов использует технику, которую иногда называют веб-скрапингом (web scraping) [Poynter, 2010]: они собирают пользовательский контент по всему интернету: с сайтов социальных сетей, блогов, микроблогов – словом, любых сайтов, где люди высказывают своё мнение или фиксируют своё поведение (см. [Schillewaert, De Ruyck, Verhaeghe, 2009]). Единицей анализа выступает дословный комментарий, а не отобранный индивид, и массив данных состоит из большого количества текстовой информации. Эти массивы данных обрабатываются с использованием естественного языка программой, способной категоризировать текст и классифицировать его в зависимости от настроения (позитивного или негативного) и в некоторых случаях – от насыщенности. В отличие от традиционных социальных обследований, аналитик, по сути, не строит связи между текстом и характеристиками респондента, который его опубликовал.

Компании всё чаще пользуются этими приёмами для отслеживания общественного восприятия продукции и услуг, а также для того, чтобы напрямую контактировать со своими покупателями, отвечая на их жалобы на той же платформе. Предпринимаются также попытки использовать эти данные для предсказывания поведения. Например, исследователи в Лаборатории HP (HP Labs) придумали технологию, использующую «Твиттер» для прогнозирования кассовых сборов от фильмов [Asur, Huberman, 2010]. В другом исследовании, основанном на анализе «Твиттера», О'Коннор и его коллеги [O'Connor et al., 2010] показали, как тесно коррелируют твиты с электоральными и социальными опросами.

Хотя, по существу, Google и не является социальным медиаресурсом, он продемонстрировал, как сбор данных поисковых запросов может использоваться для отслеживания тенденций, которые, вероятно, отражают общественное мнение и поведение. В 2008 году газета New-York Times сообщила: «Помимо привычных головных болей, кашля, жара и больного горла у гриппа появился новый симптом. Оказывается, многие болеющие американцы вводят в Google и другие поисковые системы запросы типа «симптомы гриппа» до того, как обращаются к врачам». Речь идет о Google-прогнозе заболевания гриппом (Google Flu Trends), который показывает, что число поисковых запросов о симптомах гриппа напрямую связано с данными о заболеваемости гриппом, собираемыми Центром по контролю за заболеваниями (Centers for Disease Control, CDC). Поскольку данные в Google собираются в режиме реального времени, а отчёты Центра по контролю за заболеваниями отстают по времени, поисковые запросы могут быть полезны для прогнозирования заболеваемости гриппом в разных регионах.

Крауд-методы (wisdom of crowds). Речь идет о книге Джеймса Шуровьески «Мудрость толпы: почему вместе мы умнее, чем поодиночке, и как коллективный разум влияет на бизнес, экономику, общество и государство», которая была опубликована в 2004 году. Основная её идея заключается в том, что группа людей, имеющих различные мнения и участвующих в принятии решения, лучше решит проблему или выработает прогноз, чем эксперты. Оценка каждого человека делится на два аспекта: информацию и ошибку. Величина погрешности всей группы воспринимается как нулевая. Таким

образом, использование средних показателей должно дать верную оценку возможных последствий, как, например, предсказать результаты предстоящих выборов (см., например, [The Telegraph, 2012]). Шуровьески считает, что этот метод работает лучше, если толпа разнородна и мнения собираются независимо друг от друга, но в книге недостаточно конкретики и теории, чтобы можно было судить о возможности посредством метода производить точные оценки.

Основное положение подхода коллективного разума (мудрости толпы) долгое время практиковалось на прогнозных рынках (prediction markets), подобных тем, где предсказывались результаты выборов. В качестве примера можно привести электронные рынки Айовы (Iowa Electronic Markets), созданные бизнес-школой Типпи Университета Айовы (Tippie College of Business, University of Iowa) [Intrade, 2012]. Участники прогнозных рынков обычно покупают и продают контракты, основываясь на внешних факторах, таких как экономические показатели или результаты выборов. Основная задача – предсказать результат. Поэтому вместо того чтобы использовать ответы участников, за кого они планируют голосовать, прогнозные рынки используют предсказания участников относительно победителя выборов. По сути, участники делают ставку на результат, например выиграет или проиграет президент Обама выборы 2012 года. Ротшильд [Rothschild, 2009], Эриксон и Влиезен [Erickson, Wliezen, 2008] изучили точность прогнозных рынков и не пришли к однозначному заключению.

Большие данные (big data[11]). Термин «большие данные» всё чаще используется для описания резкого увеличившегося количества данных, которые многие из нас привыкли называть административными. Под «большими

[11] AAPOR организовала рабочую группу по применению больших данных в социальных обследованиях, и в 2015 году выпущен соответствующий отчёт, который переведён на русский язык и включён в материалы ежегодной Грушинской конференции (см.: Л. Джапек, Ф. Крейтер, М. Берг. Отчёт AAPOR о больших данных: 12 февраля 2015 / Американская ассоциация исследователей общественного мнения / Пер. с англ. Д. Рогозин, А. Ипатова, Е. Вьюговская. М.: Изд-во Радуга, 2015). Поскольку настоящий текст написан несколькими годами ранее, в отношении больших данных лучше ориентироваться на новые материалы, в которых некоторые представления о больших данных скорректированы. В частности, исследователи вновь обращаются к примеру с прогнозированием Google заболевания гриппом, указывая на значительную ошибку одного из прогнозов. Исследователи отмечают, что большие данные не всегда следует рассматривать в качестве оптимального средства для построения прогностических моделей и аналитических выводов. — *Прим. перев.*

данными» подразумевается огромное количество структурированных внешних данных, которые регулярно генерируются в промышленности, бизнесе и правительстве. Раньше они часто использовались для опросов со случайным методом отбора. Например, список покупателей традиционно был основой выборки для исследования удовлетворённости покупателей. Эти же данные могут выступать в качестве вспомогательных переменных для корректировки неответов и моделирования по степени склонности.

Не так давно было замечено, что большие данные могут использоваться как альтернатива случайной выборке. Яркий тому пример – Лонгитюдное исследование динамики местоположения работодателей (Longitudinal Employer Household Dynamics, LEHD), проводимое Бюро переписи США. Это программа добровольного сотрудничества между региональными информационными биржами труда и Бюро переписи США, целью которой является получение новой информации о возможностях местных рынков труда по низкой стоимости и без дополнительной нагрузки на респондентов. Программа предоставляет государственную и региональную статистику по занятости, созданию новых рабочих мест, текучести кадров и уровню зарплат в зависимости от отрасли промышленности, возраста и пола работника. Но поскольку исходные данные основаны на записях о страховых выплатах по безработице, в них включены не все работники. Соответственно, важно понимать, что упускается из виду, – именно это позволит лучше понять те различия, которые могут появиться при сравнении с данными Бюро трудовой статистики, полученными на основании Текущего опроса населения. Программа Лонгитюдного исследования динамики местоположения работодателей несколько лет сопровождается научными исследованиями и может считаться первоклассным примером использования административных данных вместо опроса со случайным отбором. Другие случаи обращения к большим данным могут не сопровождаться анализом качества данных и ошибок, и потому в такой ситуации использование их как альтернативы случайной выборке может быть неоправданным.

Техника, которая имеет отношение к принципу работы с большими данными, – это метаанализ [Hedges, Olkin, 1985]. Одна из основных функций метаанализа состоит в том, чтобы сводить данные разных исследований

для лучшего понимания взаимосвязей переменных. Систематический обзор, например сделанный Campbell Collaboration, – вот другой вариант той же идеи. На выборах 2012 года большое внимание привлёк один вид метаанализа – *агрегирование результатов предвыборных опросов* (poll aggregation[12]). Агрегаторы довольно хорошо показали себя в отношении предсказания победителей выборов, а также процентного распределения голосов по кандидатам. Обычно они объединяли результаты ряда опросов, проведённых разными опросными компаниями, стараясь снизить *дисперсию оценки* (variance of the estimate).

Общая идея агрегирования результатов предвыборных опросов (и метаанализа) заключается в том, что больший объём выборки, который получается в результате сложения данных опросов, уменьшает дисперсию оценок. Это хорошо известный принцип, который срабатывает в большинстве случаев. Например, если мы хотим получить в результате опроса более точные оценки, увеличение размера выборки этому поспособствует. Конечно, если опрос имеет смещения из-за ошибок измерения, неответов или покрытия, точность оценок при наращивании размера выборки не увеличится, так как смещение не является функцией размера выборки (и обычно искажает дисперсию).

На последних выборах в США было несколько агрегаторов, но у всех была разная точность прогнозов, даже несмотря на то что они, вероятно, использовали одни и те же опросы. Почему? На наш взгляд, дело в том, что отличаются правила агрегирования: где-то из анализа исключались опросы, проведённые в определённой исследовательской технике, где-то – проведённые слишком давно или имеющие плохую репутацию (track records), где-то – резко отличающиеся от других. По большому счёту, агрегаторы должны были выбирать и включать в анализ только те опросы, в которых исследовалась одна и та же совокупность, что способствует агрегированию. Если они выбирали плохо, даже несколько выпадающих опросов могли сильно сместить сводные показатели.

[12] На русском языке анализом предвыборных опросов в США подробно занимался Борис Докторов. Материалы можно посмотреть на сайте фонда «Общественное мнение»: http://fom.ru/special/kto-stanet-prezidentom-ssha/list.html. – *Прим. перев.*

Схожая ситуация возникает в онлайн-исследованиях, когда стремятся снизить полученные смещения в оценках посредством агрегирования или использования разных панелей. Этот подход работает только в том случае, если панели, которые будут агрегированы, измеряют одну и ту же совокупность одним и тем же способом. Возможность выбора панелей для агрегирования ограничена, за исключением случаев, когда есть сведения, схожие с теми, что используются в агрегировании результатов предвыборных опросов.

На момент написания настоящего отчёта методы работы с большими данными (за исключением метаанализа) относятся скорее к области теории, нежели практики. Но с большой долей уверенности можно ожидать, что в скором времени эта ситуация изменится и как минимум некоторые данные, получаемые сегодня посредством опросов, будут поступать из этих источников. Тем не менее мы полагаем, что во многих случаях возникнут те же вопросы относительно неполного покрытия всей совокупности, которые стоят перед неслучайными методами отбора, равно как и вопросы различных погрешностей измерения, присущих большим данным.

3.5. ВЫВОДЫ

Цель данного раздела – обозначить два фундаментальных момента, которые будут детально рассмотрены далее. Первый заключается в том, что, в отличие от случайной выборки, неслучайная имеет много видов, каждый из которых отличается набором допущений и процедур отбора, на основании которых впоследствии будут делаться выводы о более широкой совокупности. Второй – в том, что, несмотря на это разнообразие, есть некий общий набор препятствий, которые необходимо преодолеть, если метод претендует на валидность оценок. К числу таких препятствий относятся: исключение из процесса отбора больших групп людей, частое обращение к добровольцам с недостаточным контролем за ними, высокий уровень неответов в целом (хотя эта проблема характерна и для случайных выборок).

Исследователи иногда описывают эти препятствия по-разному, и будет полезным рассмотреть парадигму, предложенную Кишем. Киш [Kish, 1987, p. 2-3] описал четыре разных класса переменных.

- *Объясняющие переменные (explanatory variables)*, которые продиктованы дизайном исследования и могут быть как независимыми, так и зависимыми переменными, между которыми и ищется связь. Это те данные, которые обусловлены анкетой.

- *Контролируемые переменные (controlled variables)* – это внешние переменные, которые могут быть с достаточной точностью проверены или в ходе отбора, или во время проведения расчётов. Типичным примером служат географические или демографические переменные, используемые либо при отборе, либо при постстратификации.

- *Возмущающие переменные (disturbing variables)* – это неконтролируемые внешние переменные, которые могут внести искажение в объясняющие переменные. Они являются неизмерёнными сопутствующими переменными (ковариатами), которые представляют интерес для измерения.

- *Случайные переменные (randomized variables)* – это неконтролируемые внешние переменные, которые рассматриваются как случайные ошибки.

Задача неслучайных методов отбора – найти любые возмущающие переменные и сделать их контролируемыми при отборе и / или расчёте оценок.

В случае с экспериментами внесение элемента случайности в наблюдения после согласования их с контрольными переменными создаст рандомизированный эксперимент, который может быть использован для построения валидных оценок причинно-следственных связей,

поскольку рандомизация гарантирует, что все возмущающие переменные будут в среднем одинаково воздействовать как на контрольную, так и на экспериментальную группы. Схожим образом случайная выборка снижает влияние возмущающих переменных, которые в неслучайных выборках с самоотбором могут привести к смещениям в оценках изучаемой совокупности. Высокий уровень неответов лишает случайные выборки этого присущего им преимущества. Иногда это поправимо — когда в основе выборки есть полные сведения о её участниках или же когда повторные контакты с неответившими дают хоть какую-то информацию о различиях между теми, кто принял участие в опросе и кто отказался.

Неслучайные методы таких преимуществ не имеют. Смещение, связанное с отбором, в большинстве неслучайных выборок создает серьёзный риск, что распределение важных сопутствующих переменных (ковариат) в выборке будет сильно отличаться от их распределения в целевой совокупности, вплоть до того, что выводы станут дезориентирующими, если не ошибочными. Чтобы быть валидными, неслучайные выборки должны основываться на некотором статистическом корректировании, которое снизит риски сильных смещений. Эффективность таких корректировок зависит от выявления важных сопутствующих переменных (ковариат), их доступности и качества. Надёжность любого неслучайного метода отбора зависит от того, насколько хорошо он решает эту фундаментальную проблему.

4 Выравнивание выборки

Выравнивание выборки (sample matching) – метод, который применяется к неслучайным выборкам уже много лет в самых разных областях знания. Его основная цель в сравнительных исследованиях – снижение смещений в оценках различий между двумя альтернативами (воздействиями или вмешательствами) посредством выравнивания выборки с контрольной группой по одной или нескольким характеристикам. Предполагается, что эти характеристики (определяемые далее как сопутствующие переменные – ковариаты) тесно связаны с объясняющими переменными (explanatary variables) и с итоговыми выборочными оценками. Для неслучайных выборок, направленных на представление некоторой большей совокупности, задача состоит в выравнивании реализуемой выборки с совокупностью таким образом, чтобы выборочные оценки стали более репрезентативными, чем без выравнивания. Основная проблема заключается в том, что неконтролируемые вспомогательные переменные (которые Киш назвал возмущающими (disturbing) переменными [Kish, 1987]) могут усиливать смещения. Выравнивание выборки направлено на преодоление этой проблемы. Основные техники выравнивания хорошо представлены в оценочных исследованиях и в анализе исследований, построенных на наблюдениях (observational studies). В последние годы выравнивание выборки применяется более широко, в том числе в маркетинговых исследованиях, опросах общественного мнения и иных социальных обследованиях.

Выравнивание выборки – интуитивно понятный и эффективный метод для снижения смещений в неслучайных выборках. Один из наиболее известных и широко распространённых методов выравнивания – квотные выборки. Их основная цель – проведение интервью с определённой группой респондентов,

которая соответствует характеристикам изучаемой совокупности. Обычно принимают во внимание наиболее доступные характеристики, такие как пол и возраст. Привлекательность подобной техники заключается в том, что «выравненная» выборка зеркально отражает изучаемую совокупность по заданным характеристикам и, возможно, это снижает смещения.

Теоретические основания выравнивания в основном развивались в исследованиях, построенных на наблюдениях (observational studies). Неслучайные выборки отбирались так, чтобы характеристики подвергаемых воздействию объектов совпали с характеристиками контрольной группы и можно было определить последствия такого воздействия. Воздействием можно назвать, например, рекламную кампанию, направленную на сокращение числа курильщиков в городе. Задача исследования в этом случае – оценить эффективность такого воздействия, которая измеряется путём сопоставления данных по городу, где кампания проводилась, с данными по похожему городу, где рекламное воздействие отсутствовало. Предполагается, что ошибка отбора будет меньше, если город, выбранный в качестве контрольного, имеет такое же распределение важных сопутствующих переменных (ковариат), как и город, где проводилась рекламная кампания.

Уже отмечалось, что большинство исследований, основанных на наблюдениях, опираются на неслучайные выборки, но иногда для оценки причинных гипотез исследователи обращаются и к случайным выборкам. Случайные выборки могут гораздо лучше репрезентировать целевую совокупность, но это не делает их пригодными для анализа причин воздействия. Например, случайная выборка взрослого населения, в которой изучается корреляция между курением и раком, недостаточна для определения причинных зависимостей между этими признаками, поскольку важные вспомогательные переменные могут быть неодинаково представлены в двух сравниваемых группах без их дополнительного явного выравнивания в процессе случайного отбора[13].

[13] В частности, в группе курящих может оказаться больше мужчин или людей старшего возраста, чем среди некурящих. В результате повышенная доля раковых больных в первой группе может объясняться как курением, так и большей подверженностью раковым заболеваниям мужчин или пожилых людей. Для выравнивания групп курящих и некурящих можно применить дополнительную стратификацию случайной выборки по полу и возрасту. – *Прим. ред.*

Соответственно, выравнивание выборки (как и взвешивание – см. раздел 6) приводит к снижению связанных с отбором смещений (selection bias), привлекая для этого внешние или вспомогательные данные. При выравнивании к вспомогательным данным обращаются в ходе отбора, при взвешивании – по его завершении. Рубин [Rubin, 1979] рекомендует комбинировать выравнивание и взвешивание для изучения каузальных эффектов в исследованиях, основанных на наблюдениях. Сегодня аналогичные подходы распространены и при решении других задач.

Выравнивание можно делать различными способами. Например, его можно проводить на индивидуальном уровне (как в нашем примере с антитабачной рекламой в городе), где каждому случаю или единице воздействия ставится в соответствие одна или несколько похожих контрольных единиц. Другой подход – частотное выравнивание распределения некоторых характеристик в контрольной выборке с их распределением в выборке, подвергаемой воздействию. Большинство квотных выборок строится на частотном выравнивании.

В теории, ошибок отбора меньше, если характеристики, используемые для выравнивания, являются ключевыми переменными, которые связаны с результатами измерений; и их сбалансированность означает, что распределения в выборке и в изучаемой совокупности совпадают. Розенбаум и Рубин [Rosenbaum, Rubin, 1983] детально описывают достижение такого баланса. Одно из неоспоримых достоинств случайного отбора состоит в том, что баланс любых сопутствующих переменных (ковариат) достигается автоматически даже тогда, когда этого нельзя добиться посредством идеального выравнивания.

В неслучайных выборках важные факторы, приводящие к смещениям, могут быть неизвестны или недоступны, поэтому сопутствующие переменные (ковариаты) останутся несбалансированными. Баланс ковариат – важнейшее условие для валидного вывода, поэтому возможность методов выравнивания выборки снижать смещения напрямую зависит от идентификации, пригодности и качества сопутствующих переменных, применяемых в качестве ковариат.

4.1. РАНДОМИЗИРОВАННЫЕ КОНТРОЛИРУЕМЫЕ ИСПЫТАНИЯ

Оценочное исследование почти всегда опирается на информацию, полученную из опросов и / или административных записей. Типичная цель оценочного исследования – установить причинную связь между воздействием и результатом (например, определённый тип изменений в доходах будет приводить к росту продовольственной безопасности). Золотой стандарт оценочного исследования – рандомизированные контролируемые испытания, которые характеризуются случайным распределением объектов между экспериментальной (несколькими экспериментальными) и контрольной группами [Shadish et al., 2001]. Экспериментальная группа подвергается воздействию, контрольная – нет.

Мы используем термин «золотой стандарт» с некоторыми уточнениями. Хотя рандомизированные контролируемые испытания обладают высокой внутренней валидностью (иногда называемой «истиной внутри исследования», truth within study), они могут не иметь высокой внешней валидности (иногда называемой «истиной вне исследования», truth beyond study). Теоретически измерение эффекта от воздействия в экспериментальной группе не испытывает влияния искажающих факторов, поскольку рандомизация автоматически создает баланс сопутствующих переменных (ковариат). Соответственно, наблюдаемые изменения в экспериментальной группе обоснованно могут быть объяснены оказываемым воздействием, а не другими причинами или побочными факторами. Однако остаётся вопрос: можно ли переносить полученные выводы за пределы конкретного экспериментального плана (экспериментальной группы)?

Обращаясь к примеру оценки социальной программы, мы можем, во-первых, определить места (или территории), в которых люди готовы к участию в оценочном исследовании. Для каждого места (территории) составляется список домохозяйств, отвечающий критериям участия в программе. Исследователь случайным образом делит список домохозяйств на две группы: половина включается в экспериментальную группу, половина – в контрольную. Рандомизация (случайное распределение по группам) обеспечивает внутреннюю валидность. Это значит, что в идеальных условиях различия,

наблюдаемые в экспериментальной и контрольной группах, связаны с программой (то есть возникают из-за воздействия программы).

Однако если мы оцениваем национальную программу, возникают потенциальные ограничения, связанные с выборкой. Отбор мест – принципиальный вопрос. Проводился ли эксперимент в одном месте или нескольких местах, и как они были отобраны? Часто при оценивании программ представители отдельных мест подают заявки на участие, из них организаторы проводят отбор – так называемая экспертная выборка (judgmental sample). Критерии отбора далеки от случайных: при отборе мест могут приниматься во внимание участие в оценке программ в прошлом, активное или пассивное согласие на участие и качество и полнота списков подходящих для эксперимента домохозяйств. Поэтому исследовательский концепт генерализации (generalizability), или внешней валидности, – это компромисс, так как потенциально важные сопутствующие переменные (ковариаты) при отборе мест не контролируются. Исследование может быть вполне валидным для отобранных мест (то есть внутренне валидным), однако возможность перенесения вывода на всю совокупность (то есть внешняя валидность) остается под вопросом.

Один из способов уменьшить угрозы для внешней валидности – повторение эксперимента в разных областях страны с разным населением. Большое количество мест может увеличить внешнюю валидность, но применение неслучайного метода отбора мест каждый раз заставляет о ней беспокоиться. Составление полного списка мест, подходящих по условиям исследования, его стратификация по ключевым переменным и формирование на его основе случайной выборки мест осуществляется редко, поскольку на практике это слишком сложно и дорого.

4.2. КВАЗИЭКСПЕРИМЕНТАЛЬНЫЙ ДИЗАЙН

Хотя рандомизированные эксперименты можно считать золотым стандартом оценочных исследований, существенные этические, правовые, законодательные и иные ограничения могут препятствовать рандомизации

в определённых исследованиях (например, отказ от лечения в клинических испытаниях). В таких случаях часто прибегают к квазиэкспериментальным планам, где не проводится случайный отбор единиц (домохозяйств или людей) в экспериментальную группу, а используются другие техники обеспечения максимально возможной внутренней валидности [Shadish et al., 2001].

Наиболее популярная техника – выравнивание выборки. Например, сначала устанавливаются места, в которых осуществляется воздействие, и затем – похожие на них места, которые включаются в контрольную группу для сопоставления. В других дизайнах выделяются домохозяйства или индивидуальные испытуемые в каком-то месте, а потом там же определяется сопоставимая контрольная выборка. Возможность провести выравнивание, например, на уровне домохозяйства – это свойство переменных, которые доступны в административных данных по определённой территории.

Смежная техника, привлекающая внимание в оценочных исследованиях, – *дизайн регрессионной дискретности* (regression discontinuity design). Идея в том, что места, не удовлетворяющие некоторому установленному критерию (критерию приемлемости, eligibility criteria) (к примеру, процент детей в школах района, получающих дотации или бесплатные обеды выше некоторого установленного порога), рассматриваются в качестве экспериментальной группы. Места, удовлетворяющие критерию, поскольку значения по нему укладываются в допустимые ограничения для данной программы, назначаются группой для сравнения. Это также форма выравнивания выборки. Одна из потенциальных угроз внешней валидности возникает тогда, когда экспериментальные и контрольные места имеют значения, слишком близкие к пороговым по отобранному критерию. В таких случаях измеряемый эффект воздействия не может быть хорошо обобщен для мест, у которых наблюдаются значительные расхождения с установленным пороговым значением.

Выравнивание выборки может быть особенно проблематичным, когда имеется большое количество сопутствующих переменных (ковариат). В литературе по оценочным исследованиям часто обсуждается одно решение для этой ситуации – моделирование по степени склонности (propensity score modelling). Цель выравнивания выборки – обеспечение

одинаковых распределений наблюдаемых сопутствующих переменных в экспериментальной и контрольной группах. В этом контексте степень склонности – это условная вероятность включения (приёма) в экспериментальную, а не в контрольную группу с учётом значений наблюдаемых сопутствующих переменных [Rosenbaum, Rubin, 1983]. Пока степень склонности неизвестна, её обычно оценивают посредством доступных вспомогательных данных, используемых в логистической регрессии. Выравнивание становится проще, поскольку используется только одна переменная – степень склонности, однако иногда наряду с ней используется ещё и другая важная ковариата. Розенбаум и Рубин [Rosenbaum, Rubin, 1983] утверждают, что выравнивание по одной только степени склонности так же снижает смещение, связанное с отбором, как и выравнивание по полному набору сопутствующих переменных. Существует мнение, что квазиэкспериментальные дизайны, в которых применяется корректировка по степени склонности (PSA), оказываются лучше рандомизированных дизайнов, однако серьёзных обоснований этого не получено [Rubin, 2008].

4.3. МЕДИЦИНСКИЕ ИССЛЕДОВАНИЯ И КЛИНИЧЕСКИЕ ИСПЫТАНИЯ

В медицинских и клинических испытаниях, помимо описанных выше рандомизированных контролируемых исследований и квазиэкспериментальных планов, используется широкий набор техник. Во многих исследованиях участвуют добровольцы. Предполагается, что эффект от новых медицинских процедур, который те оказывают на добровольцев, будет аналогичным и для населения в целом. Это весьма сильное допущение может оказаться невалидным.

Перечислим основные типы подобных исследований.

- *Рандомизированные контролируемые клинические испытания (randomized controlled clinical trials).* Субъекты рекрутируются и распределяются случайным образом на две или более

экспериментальные группы в начале исследования; результаты, наблюдаемые в группах, обычно сопоставляются по двум или более замерам через определённые промежутки времени.

- *Рандомизированные перекрёстные клинические испытания (randomized cross-over clinical trials).* Субъекты с заболеванием / медицинским диагнозом распределяются случайным образом в одну из двух экспериментальных групп в начале исследования, и после определённого периода времени, достаточного для воздействий (эксперимента) в каждой группе, вторая группа в течение такого же периода времени подвергается воздействию, которое сначала было реализовано на первой группе.

- *Рандомизированное контролируемое лабораторное исследование (randomized controlled laboratory study).* Это относится к экспериментальным работам с животными, где данные регистрируются при помощи лабораторного оборудования.

- *Когортное (инцидентное – изучающее историю заболеваний, лонгитюдное) исследование (cobort, incidence, longitudinal study).* Выбирается группа субъектов, некоторые из которых подвергались или будут подвергнуты внешнему воздействию. Субъекты изучаются посредством одного или нескольких последовательных замеров, позволяющих оценить связь между изучаемым воздействием и одним или более эффектами (результатами) от него.

Как видим, большинство медицинских исследований и клинических испытаний основаны на неслучайных выборках, в которых акцентируется внимание на внутренней валидности (посредством рандомизации в группах). Ни одно исследование не будет полезным с точки зрения внешней валидности, если не соблюдается внутренняя. Однако руководства по клиническим испытаниям уделяют много внимания статистической значимости и объёму выборки, но мало проблеме генерализируемости выводов. В итоге многие результаты этих исследований при повторных испытаниях не подтверждаются [Mayes, Horwitz, Feinstein, 1988; Young, Karr, 2011].

4.4. ВЫРАВНИВАНИЕ ВЫБОРКИ ДЛЯ МАССОВЫХ ОПРОСОВ

Цели описанных выше исследований касались исключительно производства статистического вывода об эффекте в результате воздействия или какого-либо вмешательства. Другими словами, делалась попытка ответить на вопрос, действительно ли воздействие повлекло за собой изменение, и если да, то насколько велико это изменение и куда оно направлено. В последнее время эти идеи транслируются и на выборки массовых опросов, которые куда более описательны и не нацелены на понимание специфических каузальных отношений. Поэтому выравнивание выборки в контексте массовых опросов будет несколько отличным.

Часто основной упор в изучении методов выравнивания выборки для массовых опросов делается на согласовании базовых характеристик выборки с параметрами изучаемой совокупности. С этой точки зрения изучаемая совокупность – скажем, совокупность всех домохозяйств США – аналогична экспериментальной группе в литературе по оценочным исследованиям (или случаю в исследовании методом «случай / контроль», case / control desings). Выравниваемая выборка эквивалентна контрольной группе. Как и в других видах исследований, описанных выше, основная идея заключается в том, что даже для неслучайных выборок исследователи смогут сделать статистические выводы об изучаемой совокупности, поскольку выравнивание выборки приводит к балансу сопутствующих переменных (ковриат), что уменьшает ошибки отбора, и в результате оценки исследования начинают отражать параметры совокупности.

В маркетинговом исследовании изучаемой совокупностью может быть совокупность домохозяйств, как они представлены в национальной статистике. В электоральном исследовании – население, которое, вероятнее всего, примет участие в голосовании. В этих случаях выравнивание выборки направлено на исключение или снижение ошибок отбора так, чтобы оценки выборки (к примеру, процент проголосовавших за конкретного кандидата) максимально точно отражали оценки по всей совокупности (процент голосования в день выборов за конкретного кандидата). Оценки в этой ситуации делаются исключительно по выравненной выборке. Если ошибки отбора не были исправлены,

может оказаться, что оценки также будут смещены, поскольку важные сопутствующие переменные (ковариаты) не были учтены в ходе выравнивания.

Различия между выравниванием выборки в каузальном анализе и массовом опросе важны, и их необходимо рассмотреть более детально. Во-первых, в каузальном анализе цель статистического вывода предельно ясна, и выравнивание выборки фокусируется на определении сопутствующих переменных, от которых зависят результаты или эффект экспериментального воздействия. Например, возраст, пол, генетическая предрасположенность, индекс массы тела могут рассматриваться в качестве значимых для выравнивания сопутствующих переменных при изучении воздействия диеты на снижение вероятности сердечно-сосудистых заболеваний, поскольку эти переменные влияют на результат воздействия диеты. В массовых опросах исследователь обычно заинтересован в получении большого числа разных результатов (оценок), вследствие чего используемые для выравнивания переменные (ковариаты) могут быть очень разными, в отличие от ситуации, когда опрос нацелен на получение только одной оценки. Это существенно осложняет выбор ковариат для массовых опросов[14].

Второе различие заключается в том, что результаты опросов предполагается обобщать до большой совокупности – такой, как всё взрослое население или все избиратели. Одна из техник, применяемых во многих исследованиях, основанных на наблюдении, и в медицинских исследованиях для увеличения внутренней валидности результатов, – сокращение целевой совокупности до очень специфической группы, например женщины в возрасте от 18 до 40 лет, которые никогда не имели детей, или взрослые, которые курили более 10 лет и бросили. Ограничение объекта исследования до узких подгрупп приводит к уверенности в том, что целевые подгруппы настолько

[14] Несмотря на то что в России в массовых опросах повсеместно используются сопряжённые (выравниваемые) выборки (квотные), никто не задаётся вопросом о сопрягаемых переменных. Независимо от цели и тематики исследования применяются пол и возраст, иногда добавляется образование. Причина в том, что эти демографические переменные включены в перепись и выравнивание по ним не связано с дополнительными расходами. Вопрос целесообразности и значимости этих переменных не ставится. Это пример ситуации, когда одно из наиболее значимых решений для проектирования квотной выборки принимается машинально, без какого-либо методического анализа, и, как следствие, полстеры периодически попадают в весьма затруднительные ситуации, когда их оценки оказываются полностью противоположны измеряемым реалиям. – *Прим. перев.*

близки к выборке, насколько это возможно, и, соответственно, ошибки отбора минимальны [Rosenbaum, 2005]. В большинстве опросов общественного мнения и маркетинговых исследований этот приём неприменим, поскольку требуется получить оценки по всему населению или большой и весьма разнообразной его части. Точечная оценка (например, доля поддерживающих конкретную политику или кандидата) также весьма важна для массовых опросов. В клинических испытаниях, наоборот, внимание зачастую уделяется тому, насколько сильно в результате воздействия изменились характеристики экспериментальной группы по сравнению с контрольной группой, а не определению точной пропорции тех, кто получит максимальный эффект от этого воздействия.

В настоящее время у нас отсутствуют стандарты в практиках выравнивания выборок для собираемых неслучайных, неэкспериментальных данных, которые бы позволяли распространять статистический вывод на большие совокупности. Налицо контраст с парадигмой случайного отбора. Общий подход к выравниванию состоит в определении группы базовых переменных, которые могли бы оказаться возмущающими (disturbing) или смещающими (confounding). Как и в каузальных исследованиях, эти характеристики имеют тенденцию меняться вместе с оценками, получаемыми из опроса (то есть коррелировать с ними), значения этих характеристик в неслучайной выборке могут отличаться от значений в изучаемой совокупности.

В некоторых телефонных исследованиях (особенно маркетинговых) и в большей части онлайн-исследований, как правило, применяют методы выравнивания выборки для корректировки неслучайного отбора. Например, в телефонных исследованиях могут использоваться квоты по возрасту, полу и географии, которые задают требуемое число наблюдений в каждой квотной ячейке. Эти обозначенные квотные ячейки матрицы данных отражают модель, представление исследователя о том, какой получилась бы выборка, если бы проводился валидный простой случайный отбор. К этому могут прибегать постольку, поскольку люди, ответившие и отказавшиеся от ответов, различаются. Например, в неконтролируемом телефонном опросе среди респондентов обычно обнаруживается больший процент женщин, чем в совокупности. Исследования, опирающиеся на опт-ин-панели, по той же

причине зачастую содержат квоты. Всё это относится к неслучайным дизайнам выборок, в которых исследователь пытается построить репрезентативный набор данных при помощи простых техник выравнивания выборки.

Некоторые исследователи начинают использовать более сложные методы выравнивания, аналогичные тем, которые применяются в исследованиях, построенных на наблюдениях [Rosenbaum, 2005]. Эти методы обычно опираются на более многочисленные и более разнообразные наборы сопутствующих переменных (ковариат), чем те, что применяются в каузальных исследованиях. Ниже мы опишем методы построения неслучайных выборок и покажем, что в массовых опросах они представлены и задокументированы беднее, чем в каузальных исследованиях.

4.5. ПРИМЕРЫ ВЫРАВНИВАНИЯ ВЫБОРКИ В МАССОВЫХ ОПРОСАХ

Обычно процесс начинается с определения изучаемой совокупности, которая должна быть описана в исследовании, – скажем, всё взрослое население США или избиратели. Характеристики этой совокупности собирают из разных источников. Предполагается, что источники содержат данные высокого качества, исследователи обращаются, например, к Обследованию американского общества (American Community Survey, ACS), Текущему исследованию населения (Current Population Survey, CPS), Общему социальному обследованию (General Social Survey, GSS), Американскому национальному предвыборному опросу (American National Election Survey) или к исследованию, проведённому по случайной выборке и спроектированному специально с целью получения данных для выравнивания. Следующий шаг зависит от методов, к которым обращается исследовательская организация.

Ваврек и Риверс [Vavreck, Rivers, 2008] описывают подход, основанный на принципах, которые сформулировал Риверс [Rivers, 2007]. Вначале из базы ACS отбирается случайная выборка в 38 000 человек, и она рассматривается как случаи в традиционном исследовании «случай / контроль» (case control

studies). Затем для каждой единицы наблюдения, полученной из базы ACS, находится близкий по признакам доброволец из числа составляющих опт-ин-панель. Для определения сходства между характеристиками, используемыми для выравнивания (известными как для панели, так и для сопоставляемого обследования), применяется функция расстояния (distance function). Четыре наблюдаемые переменные (возраст, раса, пол и образование), а также дополнительные переменные, измеряющие политическую приверженность и идеологические установки, определяются как ковариаты и используются для выравнивания. Далее по выравненной выборке из панели проводится опрос, и полученные в результате ответы после дополнительного статистического взвешивания, учитывающего неответы, рассматриваются в качестве оценок.

Этот подход демонстрирует пример процедуры *выравнивания один к одному* (one-to-one matching procedure), когда каждому респонденту ставится в соответствие элемент (case), или, как в описанной выше ситуации, представитель целевой совокупности взрослого населения США, случайно отобранный из Обследования американского общества. В каузальном анализе оценки каузальных эффектов сопоставляются с кейсами из контрольной группы [Rubin, 2008]. В массовом опросе каузальный эффект не оценивается, а для получения искомых распределений используются только данные из выравненной выборки (опт-ин-панель в примере Ваврека и Риверса). В результате выравнивание выборки для массового опроса не сохраняет одно из полезных свойств, которым обладают каузальные исследования, в которых процедуры выравнивания включают корреляцию между экспериментальными и контрольными значениями, что увеличивает эффективность и робастность (устойчивость) оцениваемого воздействия (экспериментального эффекта) [Rubin, Thomas, 1996].

Когда для снижения ошибок отбора требуется привлечение большого количества доступных вспомогательных переменных, выравнивание один к одному проблематично. Для точного выравнивания каждой характеристики может потребоваться очень большая панель, поскольку количество сопоставляемых ячеек возрастает геометрически с количеством выравниваемых переменных и числом их градаций. Ваврек и Риверс [Vavreck, Rivers, 2008]

решали эту проблему посредством ограничения количества выравниваемых переменных и применения меры близости вместо попыток заполнить ячейки матрицы данных.

Как мы указывали выше, стандартный подход в исследованиях, основанных на наблюдениях, когда имеется большое количество сопутствующих переменных (ковариат), – это выравнивание по степени склонности, а не по отдельным характеристикам. Такой подход сводит многомерную задачу с потенциально огромным количеством ячеек в матрице данных к более простой одномерной задаче. Поскольку степень склонности – непрерывная переменная, её можно категоризировать и затем проводить выравнивание по каждой категории или использовать меру близости для поиска наиболее точного приближения. И хотя степень склонности используют в оценочной фазе для неслучайных выборок [Terhanian et al., 2001; Lee, Valliant, 2009], для выравнивания выборки в неслучайных исследованиях её ввели только недавно [Rivers, 2007; Terhanian, Bremer, 2012].

Частотное выравнивание – другой подход к выравниванию выборки, причём в исследованиях, построенных на наблюдениях, он встречается чаще, чем выравнивание один к одному. В частотном выравнивании сначала оценивается распределение значимых переменных в изучаемой совокупности (процентное распределение по полу, возрасту и т. д.), а затем выборка конструируется таким образом, чтобы распределения переменных, по которым проводится выравнивание, были такими же, как в изучаемой совокупности. Существует множество методов частотного выравнивания: выравнивание по категориям (category matching), калиперное выравнивание (caliper matching), стратифицированная случайная выборка (stratified random sampling) или вариации парных сопряжений (см., например, [Rothman, Greenland, 1998]).

Эггерс и Дрейк [Eggers, Drake, 2011] описали одну из версий частотного выравнивания, которую они назвали выравниванием выборки, основанной на динамическом квотировании ячеек матрицы данных (dynamic quota cell-based sample matching). В этом методе в качестве основы целевых распределений берётся Общее социальное обследование (GSS). Выбор объясняется тем,

что в GSS включены разнообразные демографические, психологические и поведенческие вопросы. Исследователи делают выборку из опт-ин-панели, которая частотно выравнена с распределениями из Общего социального обследования (для этого метода не требуется выборка из GSS). Поскольку некоторые нужные для выравнивания переменные могут быть недоступны в панели, для завершения выравнивания выборки авторы рекомендуют задавать участникам панели ограниченный набор вопросов из Общего социального обследования. По полностью выравненной выборке проводится опрос, собранные ответы взвешиваются и обрабатываются для получения оценок.

Терханиан и Бремер [Terhanian, Bremer, 2012], опираясь на работу Кокрена, Тюкея и Мостеллера [Cochran, Tukey, Mosteller, 1954], а также на концепт параллельных опросов, представили иной подход. В методе используются модели по степени склонности (propensity score models), направленные на снижение смещений в неслучайной выборке, как это было изначально реализовано компанией «Харрис Интерактив» (Harris Interactive). На этапе взвешивания данных проводится выранивание распределений сопутствующих переменных (ковариат) в неслучайной выборке с распределениями, полученными в контрольной случайной выборке из той же изучаемой совокупности [Terhanian et al., 2001; Terhanian, 2008]. Основной недостаток это го подхода заключается в апостериорных корректировках, которые могут быть весьма проблематичными, если смещения в исходной выборке велики. В такой ситуации веса становятся слишком большими, существенно снижая эффективный размер выборки для получаемых оценок (effective sample size precision of estimates).

Терханиан и Бремер [Terhanian, Bremer, 2012] попытались решить эту и смежные проблемы через реализацию выравнивания непосредственно в процессе отбора, с последующим перевзвешиванием. Их методология отбора по степени склонности (propensity score select methodology) также основана на параллельных выборках (опросе по случайно сгенерированным телефонным номерам и опросе по опт-ин-панели). Оба исследования должны представлять одну и ту же целевую совокупность и опираться на одинаковую анкету или как минимум на единый набор вопросов. В дополнение к содержательным вопросам, по которым рассчитываются интересующие

исследователя оценки, оба исследования должны включать вопросы, которые идеально «учитывают все наблюдаемые и ненаблюдаемые различия» между целевой совокупностью и опт-ин-панелью. Эти сопутствующие переменные (ковариаты) могут быть демографическими, аттитьюдными или поведенческими. Исследователь также может включить некоторые эталонные вопросы (benchmark questions), которые потом могут использоваться для тестирования внешней валидности оценок. Собранные вопросы рассматриваются в качестве независимых переменных в логистической регрессионной модели, направленной на снижение различий между двумя исследованиями по параметрам, которые выбраны для улучшения внешней валидности. Модель может применяться как основа для реализации выборки в будущих исследованиях с той же целевой совокупностью и с теми же или схожими вопросами. Метод лучше всего приспособлен для больших выборок, когда первоначально проводится пилотаж с целью создания модели или для трекингового исследования.

Наконец, Гиттелман и Тримарчи [Gittelman, Trimarchi, 2009] представили подход, который назвали поиском общей средней (grand mean). Их основная цель заключалась в поддержании консистентности и надёжности от выборки к выборке, а не в увеличении внешней валидности. По одинаковой анкете они опросили по 500 респондентов в каждой из более чем 200 панелей в 35 странах. Из этих данных они составили схему сегментации, которая классифицировала людей по их покупательскому поведению, потреблению медиа, социографике в семи различных рыночных отраслях (автомобильной, потребительской электронике, банках и т. д.). Полученные сегменты могут использоваться для классификации будущих респондентов из практически любой выборки, что позволит обеспечить правильные распределения выборки по сегментам для изучаемых рынков. В дальнейшем они начали включать части случайных телефонных опросов (RDD) в выборки для совершенствования сегментов [Gittelman, Trimarchi, 2010]. Так они надеются увеличить репрезентативность панельных выборок. Гиттелман и Тримарчи продолжают развивать свой подход, и одна из важнейших составляющих этого развития – раскрытие всех допущений, на которых он основывается.

4.6. ВЫВОДЫ

Обсуждавшиеся выше техники выравнивания выборок – весьма интересные и инновационные приложения для массовых опросов. Они опираются на методы, которые применяются в других областях знания уже многие годы. Хотя наше обсуждение этих методов в основном не касалось вопросов валидности, что подтверждало бы их эффективность, такие исследования существуют, и мы рекомендуем читателям ознакомиться с ними, прежде чем выносить суждения на этот счёт. Мы также отмечаем, что есть некоторые ключевые предположения, лежащие в основе всех описанных методов, и их необходимо учитывать в работе с результатами исследований, основанных на выравнивании выборок.

Во-первых, данные, используемые для контрольной группы, должны быть высокого качества и содержать минимум ошибок. Все три описанных выше метода опираются на опросные данные, и оценка ошибки в этих контрольных опросах чрезвычайно важна.

Во-вторых, следует иметь в виду, что эти техники изначально разработаны для оценки единичных экспериментальных эффектов. Обследования, напротив, проводятся для оценки многих факторов. В каузальных исследованиях эффекты от воздействия оцениваются посредством данных как из подвергаемых воздействию кейсов, так и из сопряженной выборки (эффекты рассчитываются как разница между этими двумя группами). В опросах для производства оценок используются только данные выравненной выборки. Робастность (устойчивость) и надёжность выводов, которые опираются только на выравненную выборку, изучены недостаточно.

В-третьих, возможно, наиболее важно, что опросы обычно охватывают большое количество тем, и даже единичный опрос зачастую направлен на производство множества оценок. Различные темы имеют разные сопутствующие переменные (ковариаты). Не существует универсального набора сопутствующих переменных, которые можно применять для корректировки всех смещений во всех возможных темах, которым посвящены исследования. Количество сопутствующих переменных, необходимых для одного

исследования, может быть весьма большим. Как мы отмечали выше, главное предположение для любого неслучайного метода отбора заключается в том, что ключевые сопутствующие переменные определены, измерены и сбалансированы. Если это предположение не выполняется, выборка может содержать серьезные смещения, связанные с отбором.

Как отмечает Розенбаум [Rosenbaum, 2005], «поскольку даже идеально спроектированный дизайн неэкспериментального исследования будет содержать ошибки и неопредёленности, проведения единичного исследования зачастую недостаточно, и требуются повторные замеры. В повторных исследованиях следует стремиться воспроизвести действительные экспериментальные эффекты, если они проявляются, без воспроизводства каких-либо смещений, которые могут влиять на результаты основного исследования». Этот мудрый совет может быть непрактичным для опросов, где негативные последствия от небольших ошибок в оценках могут быть менее значимыми, чем ошибки при оценивании эффектов от воздействия. Однако для опросов, использующих выравнивание выборки, учитывать этот совет при рассмотрении методов повторных замеров вполне целесообразно. Например, если проводится несколько таких исследований, некоторые вопросы могут быть включены в каждый замер, и тогда воспроизводимость не потребует значительных расходов.

В других областях знания, где применяются сопоставимые исследования (matched studies), необходимо осознавать и принимать во внимание тот факт, что результаты могут быть неустойчивыми и иногда приводить к выводам, которые впоследствии будут противоречить друг другу. Например, Мейс, Хорвиц и Файнштайн [Mayes, Horwitz, Feinstein, 1988] описали 56 тем, изученных в контролируемых экспериментальных планах (case-control studies), результаты которых противоречили последующим исследованиям. Они отмечают, что «противоречия могут возникать как вследствие того, что изучаемые в исследовании причинные связи не основаны на рандомизированных экспериментальных планах, так и оттого, что группы организуются и данные собираются без привлечения стратегий, позволяющих избежать или уменьшить смещения». Неслучайные выборки, основанные на методах выравнивания, сталкиваются с теми же проблемами.

Наконец, важнейший вопрос, который мы не затронули, – это транспарентность описываемых методов. Транспарентность – важнейшая тема в исследованиях, основанных на наблюдениях. Она в той же мере важна и для массовых опросов, основанных на выравнивании выборки. Ванденброуке с коллегами [Vandenbroucke et al., 2007] описывают деятельность группы методологов, исследователей и редакторов, которые составляли рекомендации по улучшению качества отчётности в исследованиях, основанных на наблюдениях. Они называют эти рекомендации «Улучшение отчётности исследований, основанных на наблюдениях, в области эпидемиологии» (Strengthening the Reporting of Observational Studies in Epidemiology, STROBE). Это дает нам концептуальную схему, которая может рассматриваться в качестве модели для разработки требований к отчётности в массовых опросах по неслучайным выборкам в целом и по выравненным выборкам в частности.

5 Сетевая выборка (network sampling)

Представьте, что вы изучаете мужчин, которые имели секс с другими мужчинами (далее по тексту – МСМ) в маленьком городе. Вы предполагаете, что число таких мужчин составит 2 % от всего мужского населения. Каким образом вы будете их отбирать? Можно использовать стандартную основу выборки для всего населения, то есть прежде чем начать опрос, проверять каждого потенциального респондента на его соответствие скринингу. Вам придется найти и проинтервьюировать в 50 раз больше людей, чем нужно для опроса. Скрининг должен быть довольно долгим, чтобы выстроились доверительные отношения и респондент мог раскрыть свой МСМ-статус. Такой подход будет материально затратным и в целом неприемлемым. Другой подход – найти и отобрать МСМ там, где они собираются, например в барах, парках, на публичных мероприятиях, и опросить респондентов там. Но такие выборки, скорее всего, не затронут многие подгруппы МСМ в городе.

Сетевая выборка предлагает альтернативу. Поскольку МСМ чаще всего являются социально взаимосвязанной группой, в построении выборки могут помочь их контакты в социальных сетях. Если исследователь найдёт даже небольшое число МСМ и установит с ними достаточно доверительные отношения, можно будет получить контакты их знакомых, которые, в свою очередь, подключат своих знакомых МСМ, и т. д., пока не будет достигнут необходимый размер выборки. Такой подход доказал свою эффективность в формировании больших и разнообразных выборок во многих так называемых труднодоступных группах, где традиционные методы отбора оказываются неэффективны. Строить статистические выводы по результатам исследований с такой выборкой сложно, поскольку большинство из них не опирается на случайную выборку с известными вероятностями отбора.

То, что описано выше, – пример сетевой выборки с отслеживанием связей (link-tracing network sampling), стратегии использования сетевых связей для построения выборки. Отличительная черта такого подхода заключается в том, что последующие участники выборки отбираются из числа сетевых контактов предыдущих. Таким образом, для увеличения выборки выявляются сетевые связи, а структура сети используется, чтобы упростить поиск новых участников. Быстрое распространение социальных сетей в интернете – таких, например, как Facebook, – дало возможность строить сетевую выборку из очень большого числа людей, и это удобный и недорогой способ получить к ним доступ.

Сетевая выборка, по большому счёту, не является подходом, опирающимся только на неслучайные выборки. По факту, в первое время в научных работах по статистике (например, [Goodman, 1961; Frank, 1971; Thompson, 1992]) её основы связывались с методами построения случайных выборок. В современных исследованиях сетевая выборка, основанная на подключении связей, доказала свою целесообразность, когда неприменимы строгие допущения, необходимые для случайных методов отбора. Например, в случае с редкими этническими меньшинствами (например, [Welsh, 1975; Snow et al., 1981]), людьми в группе риска болезней типа ВИЧ (например, [Klovdahl et al., 1994]) или социально отчужденными работниками [Bernhardt et al., 2009].

5.1. МОТИВАЦИЯ

Выборки с отслеживанием связей внутри сети чаще всего используются для труднодоступных групп, когда традиционные методы отбора не работают или оказываются нецелесообразными. Их также используют для снижения затрат в случаях, когда традиционные методы доступны, но дорогостоящи. Несмотря на то что сетевая выборка может снизить стоимость сбора данных, для достижения требуемого уровня точности может потребоваться значительно большее число выборок, поскольку данные, как правило, зависят от выбранной сети. Следовательно, не всегда ясно, действительно ли стоимость информации при такой стратегии ниже, чем при традиционных подходах.

Для некоторых групп населения стандартные методы могут быть физически неосуществимы. Например, сексуальные меньшинства могут быть стигматизированы в обществе, и потому идентификация представителей изучаемой совокупности в доступных основах выборки может быть затруднительной (например, [Zea, 2010]). Какие-то группы людей, например, небольшие этнические меньшинства, встречаются довольно редко, потому выборки, построенные на доступных основах, смогут затронуть только очень малое число их представителей, и на это уйдет много времени (например, [Kogan et al., 2011]).

Киш и Калтон [Kish, 1965, 1987; Kalton, 1993, 2003, 2009] рассматривали различные техники случайного отбора для малочисленных групп людей, позволяющие проводить валидные, основанные на проектировании оценки характеристик интересующих нас редко встречающихся групп населения. Техники случайного отбора для таких групп включают:

- создание списков (основы выборки) малочисленных групп населения;
- многократный отбор (multiplicity sampling);
- непропорциональную стратифицированную выборку (disproportionate stratified sampling);
- непересекающиеся многоосновные выборки (non-overlapping multiple frame designs);
- пересекающиеся многоосновные выборки (overlapping multiple frame designs);
- накопление соответствующих элементов выборки из предыдущих опросов разных слоев населения;
- дизайны выборки, включающие скрининг для малочисленных групп населения (sample designs involving screening for the rare subpopulation);
- двухступенчатую выборку (two-phase sampling).

Эти техники используются многие годы. Национальный опрос об иммунизации, проведённый Центром по контролю и профилактике заболеваний в США (CDC, Centers for Disease Control and Prevention, 2005), является примером скринингового опроса домохозяйств, в которых есть дети

в возрасте от 19 до 35 месяцев. В качестве недавнего примера использования многоосновной выборки может выступить опрос американских мусульман (Исследовательский центр Пью, Pew Research Center) 2011 года. При использовании случайного набора номера (RDD) только 0,5 % респондентов идентифицировали себя как мусульмане (или американские мусульмане), потому принцип построения выборки данного исследования основывался на комбинировании случайного набора номеров стационарных и мобильных телефонов в районах с высокой концентрацией мусульманского населения и включал повторное обращение к респондентам-мусульманам, отобранным посредством случайного набора номера в других исследованиях.

Вместе с тем в некоторых ситуациях использование случайных методов отбора для малочисленных групп оказывается практически нецелесообразным, поскольку для набора желаемого количества необходимых интервью такие техники, как скрининг представителей малочисленных групп населения, обходятся слишком дорого. Стоимость может быть непомерно высокой, если представители малочисленных групп населения предположительно встречаются в 1 % или менее домохозяйств региона, где проходит опрос.

В других ситуациях техники случайного отбора не могут применяться ввиду определённых характеристик, присущих малочисленным группам. Например, не подходят техники случайного отбора, основанные на выборке домохозяйств, если значительная часть подгруппы находится в приютах или является бездомной. В таком случае создать список представителей малочисленной группы населения невозможно. То же самое касается малочисленных групп, которые являются «скрытой совокупностью», например тех, кто практикует противоправное поведение, скажем, потребление инъекционных наркотиков.

Другие неслучайные методы отбора для такого населения применимы. Три известные альтернативы – это квотная выборка, целевая выборка и выборка по времени и месту (time-location sampling). И квотирование, и целенаправленный отбор являются полностью неслучайными методами.

Целевая выборка [Watters and Biernacki, 1989] – это неслучайный метод отбора, который сочетает этнографическое картографирование с квотами для отбора, квотирование по времени и месту, а также экспертную сетевую выборку (peer-referrals constituting network sampling). Это точный, прагматичный и всё же основанный на неслучайном отборе подход, созданный для сбора репрезентативных данных о труднодоступных группах населения. Он используется с определённым успехом [Carlson et al., 1994]. Робинсон и его коллеги [Robinson et al., 2006] сравнили целевую выборку и выборку, управляемую респондентами (вариант с сетевой управляемой выборкой подробно описан ниже). Они установили, что качество выборок, построенных при помощи этих двух методов, сопоставимо, но для целенаправленного отбора требуются значительно большие усилия исследователей, а для выборки, управляемой респондентами (RDS), – гораздо большее финансовое стимулирование респондентов.

Выборка по времени и месту (time-location sampling) [Muhib et al., 2001; MacKellar et al., 2007] может опираться на случайную выборку с известной основой, но эта основа имеет двухуровневую структуру. Поскольку взаимосвязь двухуровневой основы и целевой группы неоднозначна, этот метод отбора по своей сути является неслучайным. Верхний уровень основы выборки состоит из списка мест, где собираются представители изучаемой совокупности, с учётом времени, когда это происходит. Такие комбинации места и времени трактуются как страты. В каждой из страт представители совокупности отбираются при помощи метода, который указали исследователи, чаще всего это перепись или стратегия случайного отбора. Такой подход технически является стратегией случайного отбора, так как вероятность отбора всех представителей может быть вычислена исходя из их принадлежности к определённым стратам. На практике же использование этого метода сильно ограничено ввиду того, что тяжело разбить всю изучаемую совокупность на страты. В случае с MSM, например, выбранные комбинации места и времени могут покрыть большинство центров гей-культуры, но будут упущены из виду те сегменты представителей MSM, которые нечасто такие места посещают. Выборка по времени и месту очень похожа на перехватывающий отбор (location-intercept sampling), при котором интервьюеры стоят в намеченных местах и систематически отбирают проходящих мимо людей. Такой отбор

также можно рассматривать как разновидность кластерной выборки, где кластерами служат комбинации места и времени. Как и в кластерной выборке, здесь часто можно построить случайную выборку внутри каждого кластера, но взаимосвязь кластеров со всей изучаемой совокупностью остается неясной.

Преимущество сетевой выборки состоит в том, что социальные связи респондентов позволяют распространить основу выборки за пределы видимых или доступных представителей изучаемой совокупности таким способом, который, надо надеяться, позволяет относиться к полученной выборке как к (более или менее) случайной.

В Бразилии, например, Киндал и его коллеги [Kendall et al., 2008] выявили, что сетевая выборка, управляемая респондентами (RDS), оказалась менее дорогостоящей и более разнородной для отбора мужчин, которые имели секс с другими мужчинами, чем выборка по месту и времени их сбора. И это часто характерно для малочисленных, стигматизированных или живущих вне домохозяйств групп населения. Ввиду взаимозависимостей между последовательными выборками (dependencies between successive samples) сетевые выборки чаще всего не являются первоначальным выбором исследователей общественного мнения. Тем не менее во многих случаях они представляют собой наиболее теоретически обоснованную и жизнеспособную альтернативу.

5.2. ИСТОРИЯ СТАТИСТИЧЕСКИХ ПУБЛИКАЦИЙ

Хотя выборка, построенная на отслеживании внутрисетевых связей, появилась даже до исследований Коулмана ([Coleman, 1953]; см. также [Handcock, Gile, 2011]), её всё равно принято связывать с работой Гудмана 1961 года, в которой он представил вариант выборки, основанной на внутрисетевых связях, который описал как «выборка методом снежного кома, имеющая *s* волн *k* контактов» (*s* stage *k* name snowball sampling). Эта исходная вероятностная формула предполагает, что существует полная основа выборки, из которой получена случайная выборка первоисточников, или стартовых точек (seeds). Заданное количество контактов (*k*) каждого источника

фиксируется на первом этапе, или волне снежного кома. Контакты (*k*) каждого респондента первой волны формируют вторую волну и т. д., пока не наберётся заданное количество волн (*s*). Гудман призывает к использованию такой выборочной стратегии для получения статистических выводов о количестве сетевых структур разного типа во всей сети. Так, особое внимание он обращает на циклы в сети — многоугольные (многореберные) циклы, в том числе треугольные, квадратные и т. д. В противовес недостаткам выборок с отслеживанием сетевых связей, которые применялись в то время, метод снежного кома, по мнению Гудмана, увеличивает эффективность выборки. В частности, если в выборке методом снежного кома использовать заданное число волн (*s*) для оценки числа *s*-угольных циклов (например, трёх волн для треугольных, четырех волн для квадратных), то для достижения желаемой точности в оценках числа *s*-угольных циклов потребуется обследовать значительно меньшее число узлов, чем потребовалось бы при их простом случайном отборе.

Можно также привести в пример работы Оува Франка начала 1970-х годов (например, [Frank, 1971], краткое изложение в [Frank, 2005]), основанные на формуле Гудмана и расширившие типы выборок, объектов исследований и алгоритмов оценивания, для которых могут применяться выборки с отслеживанием внутрисетевых связей. Как и Гудман, Франк рассматривает случаи, в которых первоисточники (seeds) могут быть отобраны из изучаемой совокупности с использованием случайной выборки и где могут быть выявлены все необходимые для вычисления вероятности отбора характеристики сети. В подходах, основанных на проектировании выборки, для статистических выводов необходимо знать вероятности отбора. Хендкок и Гайл [Handcock and Gile, 2011] описали, какие ограничения эти условия накладывают на сетевые выборки с отслеживанием связей и на статистические выводы, основанные на дизайне, и пришли к выводу, что многие из этих условий не имеют разумного обоснования в ситуациях, встречающихся на практике.

Многократный отбор (multiplicity sampling) (ввели Бирнбаум и Сиркин в 1965 году [Birnbaum and Sirken, 1965]) – это особая разновидность дизайна сетевой выборки с отслеживанием связей, при котором известна вероятность отбора. Здесь основа выборки конструируется таким образом, что

некоторые респонденты могут быть отобраны более одного раза. Например, в исследовании Брика [Brick, 1990] молодёжь имела шансы дважды попасть в выборку, так как в телефонном опросе спрашивали обо всех молодых людях – как о проживающих в этом домохозяйстве, так и о тех, у которых в этом домохозяйстве живёт только мать. Этот подход также использовался для увеличения возможности попадания в выборку (over-sample) людей с редкими заболеваниями или из определённых этнических меньшинств (подробнее см. [Kalton and Anderson, 1986]). Многократный отбор основан только на одной ступени или волне сетевой выборки и зависит от чётко обозначенных взаимосвязей – как, например, родственные узы или географическое соседство, – которые позволяют корректно определить вероятность отбора (с учётом вероятности повторного отбора).

Работы Стива Томпсона и его коллег (например, [Thompson, 1990; 2006a; 2006b; Felix-Medina and Thompson, 2004]) сместили многоволновые случайные выборки от дизайнов отслеживания связей в сферу практической выборки и анализа. В частности, Томпсон и Себер [Thompson, 1992; Thompson and Seber, 1996] подчёркивали адаптивную природу многих дизайнов с отслеживанием связей. Традиционно дизайн выборки полностью определяется до сбора данных, включая вероятность отбора каждой из единиц на каждой ступени. В адаптивном дизайне информация, собранная во время исследования, может повлиять на последующий сбор данных. Сетевые выборки чаще всего такие: обнаруженные в ходе отбора связи сети с предыдущими респондентами могут повлиять на расчёт вероятности отбора. Для валидных статистических выводов, полученных с использованием доступных методов, всё же нужно, чтобы процесс отбора зависел только от характеристик сети, которые можно зафиксировать. Удивительно, но выборки, построенные на внутрисетевых связях, начало которым положили случайные выборки, часто отвечают этим критериям.

Томпсон и его коллеги [Thompson and Frank, 2000; Thompson, 2006a,b] предложили множество методов использования адаптивных выборочных стратегий для построения выборок и производства статистических выводов. Основное ограничение этих методов заключается в том, что им требуется наличие и исходной случайной выборки (initial probability sample),

и основы выборки, включающей всю изучаемую совокупность. Исключение составляют работы Феликс-Медины и его коллег (в первую очередь, Феликс-Медина и Томпсон [Felix-Medina and Thompson, 2004]), в которых требуется частичная основа выборки судебных округов. На практике такая основа зачастую отсутствует, и это является стимулом к использованию стратегий сетевой выборки.

Особо следует отметить работу Томпсона и Франка [Thompson and Frank, 2000], где представлен возможный способ статистического вывода из выборок, основанных на внутрисетевых связях, который противопоставляется прежней концепции, основанной на проектировании. Статистические модели могут подходить для сетевой выборки с отслеживанием связей без моделирования процесса отбора и без обязательного знания вероятности отбора каждой единицы наблюдения, если выборка зависит только от обследованной (наблюдаемой) части сети. Опять же, этот подход требует наличия исходной случайной выборки.

5.3. НЕСЛУЧАЙНЫЕ СЕТЕВЫЕ ВЫБОРКИ

Фраза «выборка методом снежного кома» часто означает совсем не то, что подразумевает приведённая выше вероятностная формулировка. Иногда так называют конформную выборку (см. раздел 3), стартовавшую с неслучайной выборки и далее расширенную посредством привлечения контактов (обычно всех) каждого предыдущего участника. Такая выборка не является случайной, поскольку вероятность отбора определяется сначала исходной неслучайной выборкой, а потом уже связями с предыдущей конформной выборкой (см., например, [Biernacki, Waldorf, 1981; Handcock, Gile, 2011]).

Примеры использования выборок методом снежного кома в неслучайном варианте разнообразны: от Троу [Trow, 1957], изучавшего радикальных политиков в Беннингтоне (США, штат Вермонт), и Каплана и его коллег [Kaplan et al., 1987], изучавших потребителей героина в Нидерландах, до МакКензи и Мистиена [McKenzie, Mistiaen, 2009], изучавших потомков японцев в Бразилии.

5.4. СЕТЕВАЯ ОНЛАЙН-ВЫБОРКА (ONLINE NETWORK SAMPLING)

Говоря о социальных сетях, нельзя обойти вниманием интернет – как глобальную сеть в целом, так и отдельные сети, например Facebook и Twitter. Бесспорно, эти и другие онлайн-форумы для общения предоставляют возможности строить выборки. Впервые онлайн-выборка по социальным сетям появилась в литературе по электронной инженерии и вычислительной технике – там описывались программные «поисковые роботы» («черви» – crawlers), придерживающиеся определённых алгоритмов для отслеживания интернет-связей (например, дружба на Facebook) от одного пользователя к другому, собирая доступные данные из выбранных профилей пользователей. Используемые поисковые алгоритмы различаются и являются объектом изучения во многих исследованиях. Например, Гьйорка и его коллеги [Gjoka et al., 2011] сравнили несколько простых подходов, а также представили более детальный поисковый механизм, стараясь приблизиться к репрезентативным выборкам из социальных сетей. Другие специализированные алгоритмы были направлены на иные перспективные характеристики сетей. Например, пограничная выборка (frontier sampling) преследовала цель вычислить, насколько вероятно, что некоторые части сети могут быть не связаны [Ribeiro, Towsley, 2010]. В такого рода подходах, тем не менее, и сейчас пользователей не просят заполнить новый опросник, а просто собирают доступную в онлайне информацию.

Использование социальных онлайн-сетей для построения выборки в социальных обследованиях пока недостаточно изучено. Вайнарт и Хекаторн [Wejnert, Heckathorn, 2007] представили версию управляемой респондентами выборки (RDS), реализуемой через интернет (описана ниже), которую они называют WebRDS. Респонденты рекрутировали своих сетевых знакомых по электронной почте, и анкеты заполнялись онлайн. Этот метод был опробован в исследовании студентов колледжа и оказался эффективным для рекрутирования большого числа студентов за короткое время. Некоторые характеристики исследуемой совокупности воспроизводились довольно точно, но в некоторых оценках были выявлены существенные смещения, связанные с разницей в использовании электронной почты различными подгруппами.

5.5. ВЫБОРКА, УПРАВЛЯЕМАЯ РЕСПОНДЕНТАМИ: АППРОКСИМАЦИЯ СЛУЧАЙНОЙ ВЫБОРКИ

Выборка, управляемая респондентами (respondent-driven sampling) [Heckathorn, 1997], – это такой подход к построению выборки и анализу, который находится где-то между практичной конформной выборкой и случайной выборкой, предоставляющей возможности построения статистических выводов. Такой, казалось бы, малореальный подход строится на двух вещах: хорошо продуманном дизайне выборки и разумной доли допущений, пусть некоторые из них фактически не тестируемы.

Метод получил своё название из-за раздачи купонов, которой управляют сами респонденты. Дизайн RDS-выборки включает два ключевых нововведения, оба относятся к рекрутированию при помощи купонов. В большей части дизайнов выборок, основанных на отслеживании связей, респондентов просят перечислить всех знакомых из изучаемой совокупности, а исследователи уже отбирают из них тех, кто войдёт в выборку, и здесь могут возникнуть проблемы с конфиденциальностью при исследовании стигматизированных групп населения. В RDS же респондентам выдают купоны с уникальными номерами для передачи их знакомым, которые таким образом получают доступ к участию в исследовании. Подобное решение значительно снижает беспокойство, связанное с конфиденциальностью, и делает возможным отбор среди намного более широкого круга людей.

Обычно в выборках с отслеживанием связей реализуется как можно большее число контактов каждого участника. И второе серьёзное нововведение – ограничение числа респондентов, которые могут быть рекрутированы. Это нововведение позволяет RDS-выборкам заданного размера пройти от стартовых точек выборки (seeds) по более длинным цепочкам связей (сделать больше волн), чем в других выборках с отслеживанием связей. Для сравнения: если респондент предоставляет в среднем 5 контактов для исследования, то выборка от одной стартовой точки (seed) наберет 150 респондентов за три волны. Если же в среднем от одного респондента берется два контакта, то для получения 150 респондентов от одного источника потребуется более шести волн. Большее количество волн означает, что итоговая выборка меньше зависит

от исходной выборки (initial sample). Во многих методах анализа это снижение зависимости итоговой выборки от исходной позволяет расценивать полученные результаты как аналогичные результатам случайной выборки.

Очевидно, чтобы считать, что такие данные получены посредством случайной выборки, необходимо сделать несколько допущений. Это зависит от алгоритма оценивания, и сейчас вводятся новые алгоритмы, которые меняют набор необходимых допущений (обычно заменяя один набор допущений другим). Здесь мы опишем допущения алгоритма, введённого Вольцем и Хекаторном [Volz, Heckathorn, 2008]. Известный как алгоритм RDS-II, *или VH*, он также тесно связан со стандартными статистическими методами, в особенности теми, которые описывали Хансен и Хурвитц [Hansen, Hurwitz, 1943]. Более подробно об этом пишут Гайл и Хендкок [Gile, Handcock, 2010].

Один набор допущений необходим для того, чтобы убрать зависимость от исходной выборки или стартовых точек (seeds). Предположим, представители совокупности в изучаемом городе имеют связи только с людьми из своего района. Тогда вне зависимости от числа волн выборка, начало которой находится в одном районе, никогда не выйдет за его пределы. Таким образом, районный состав выборки будет полностью зависеть от исходной выборки. Это крайний случай гемофильности (homophily) – тенденции, когда люди притягиваются к себе подобным чаще, чем к другим людям [McPherson et al., 2001]. Даже относительно слабые формы гемофильности могут привести к тому, что выборка RDS будет слишком сильно зависеть от состава начальной выборки. Иногда для преодоления этого эффекта требуется много волн. Кроме того, если есть какие-то части сети, совершенно недостижимые из других её частей, то невозможно от выборки, начатой в одной подгруппе, перейти в другую. Это называется «несвязанный граф» (disconnected graph). Алгоритм VH предполагает, что граф связанный и что гемофильность достаточно низка для числа волн в выборке.

Поскольку статистические выводы в данном случае зависят от дизайна исследования, такой подход также требует известной вероятности отбора всех единиц наблюдения. Алгоритм VH предполагает, что эта вероятность прямо пропорциональна рангу респондента (respondent's degree) или числу его

контактов в изучаемой совокупности. На интуитивном уровне это имеет смысл, поскольку индивиды с большим числом контактов имеют большую вероятность быть отобранными. Это основано на моделировании процесса отбора как цепи Маркова в пространстве представителей совокупности (описано более подробно у Гайла и Хендкока [Gile, Handcock, 2010]). Если бы процесс был настоящей цепью Маркова, то каждый респондент использовал бы для рекрутирования только один социальный контакт, выбранный совершенно случайно из числа всех своих контактов. Ещё один важный момент связан с допущением, что людям не запрещён повторный отбор предыдущих участников опроса. На практике отбор проводится без возвращений (without-replacement), то есть человек, который уже был отобран, не может повторно участвовать в отборе. Следовательно, для модели, основанной на цепи Маркова, требуется, чтобы размер совокупности был значительно больше размера выборки. Если все связи взаимны, то есть Джо является контактом Сью тогда и только тогда, когда Сью является контактом Джо, то в этой модели вероятность отбора респондентов прямо пропорциональна числу их контактов. Оценка такой вероятности отбора также требует точного измерения числа контактов респондента.

Введены и новые алгоритмы, позволяющие ослабить некоторые из этих допущений. Алгоритм, описанный Гайлом [Gile, 2011], сразу предполагает выборку без возвращений (without-replacement), что снимает требование большого размера совокупности. Тем не менее, поскольку он анализирует потенциально конечную совокупность, ему требуется оценка размера этой совокупности. В алгоритме, разработанном Гайлом и Хенкоком [Gile, Handcock, 2011], учтены допущения выборки без возвращения; он предлагает подход к снижению зависимости от стартовых точек (seeds), а также смягчает требования низкой гемофильности или достаточного числа волн выборки. Новый подход делает эти допущения менее строгими, но требует большей степени доверия к получаемой из выборки информации для оценки уровня гемофильности.

Хотя эти алгоритмы могут обеспечить относительно несмещённые точечные оценки с указанными допущениями, возможность вычисления дисперсии оценок всё же сомнительна. Поскольку RDS создаёт зависимые выборки,

каждая новая волна выборки добавляет значительно меньше информации, чем если бы это была простая случайная выборка. По этой причине дисперсия итоговых оценок существенно больше, чем при простой случайной выборке того же размера. Этот феномен усугубляется при высокой гемофильности, которая приводит к росту зависимости между элементами выборки. В работах Салганика [Salganik, 2006] и Гоеля [Goel and Salganik, 2009, 2010] эти вопросы рассмотрены более подробно.

Оценка погрешности RDS-выборок также довольно проблематична. Во всех работах (Салганик [Salganik, 2006], Фольц и Хекаторн [Volz, Heckathorn, 2008], Гайл [Gile, 2011], Гайл и Хендкок [Gile, Handcock, 2011]) авторы описывают алгоритмы расчёта стандартных ошибок для RDS-оценок. И хотя каждый из них может быть полезен при оценке погрешности, нигде нет описания всех потенциальных источников расхождений в сложном процессе отбора. Дальнейшие размышления об оценке погрешности неслучайных выборок представлены в разделе 7.

5.6. ВЫВОДЫ

Сетевая выборка вызывает одновременно и вопросы, и интерес. С одной стороны, чтобы выводы считались валидными, может потребоваться большое число часто не тестируемых предположений. Из-за взаимозависимости единиц наблюдения расхождения в итоговых статистических оценках могут быть велики. По этим причинам сетевая выборка может не быть приоритетной при выборе дизайна исследования.

С другой стороны, сетевая выборка предлагает практически выполнимый подход к некоторым проблемам отбора, не решаемым другими методами. Когда не существует валидной основы выборки, необходимой для традиционных методов отбора, исследование социальных взаимосвязей между известными и неизвестными представителями изучаемой совокупности может быть единственным способом изучить эту совокупность. Если совокупность изнутри социально взаимосвязана, сетевая выборка может

быть самым точным из доступных методов. Малочисленные группы, выделенные на основе иммиграционного статуса, этнической принадлежности, сексуальных практик, использования наркотиков или типа занятости, могут быть охвачены при помощи сетевой выборки. В таких случаях исследователям остаётся только быть предельно внимательными при разработке стратегии отбора и анализа, которая опирается на обоснованные предположения и допущения, что позволит достичь максимально возможных при данных условиях научно-теоретических и практических результатов.

6 Методы оценки и взвешивания

Исследователи обращаются к выборкам, когда сбор данных во всей совокупности (например, перепись) весьма затруднителен. В зависимости от цели исследования использование выборки может потребовать некоторой процедуры оценивания, чтобы по ответам, полученным в выборке, сделать верные выводы о совокупности.

В целом задачи исследования могут быть разделены на две категории. Первая категория – исследования для получения статистических показателей, имеющих отношение только к выборке. Примеры включают когнитивное тестирование или пилотажные исследования, в которых статистические показатели выборки задают набор метрик для оценки опросных процедур. Цель может состоять в том, чтобы распространить полученные оценки на значительно большую новую выборку, но исследователи не пытаются сделать статистические выводы обо всей изучаемой совокупности. Как уже обсуждалось в разделе 4, цель может быть и другой — оценка внутренней валидности некоторых измерений. Хотя для этих целей используются в равной мере случайные и неслучайные выборки, большинство обращается к неслучайным, поскольку они позволяют существенно сократить затраты на сбор данных.

Вторая категория – исследования, направленные на перенос выводов с выборки на изучаемую совокупность посредством производства оценок. Именно эта категория обсуждается в настоящем разделе.

Начнём с концепта «изучаемая совокупность» (target population[15]). Лор [Lohr, 1999] определяет изучаемую совокупность как «полное множество наблюдений» в исследовании, заданное дизайном опроса. Используя одну или несколько основ выборки, случайная выборка связывает с каждым индивидом изучаемой совокупности вероятность его включения в выборку. В парадигме основанного на проектировании статистического вывода (design-based inference paradigm) эти вероятности отбора позволяют исследователю рассчитать оценки для изучаемой совокупности. Напротив, в неслучайных дизайнах основа выборки часто отсутствует, поэтому связь с изучаемой совокупностью не может быть определена однозначно, а вероятности отбора и вовсе не определены. В результате точные основанные на дизайне выборочные веса, обратно пропорциональные вероятностям отбора, не могут быть рассчитаны, и, следовательно, этот метод статистического вывода невозможен для неслучайных выборок. В этой связи некоторые исследователи утверждают, что недопустимо говорить о статистических свойствах оценок для неслучайных выборок [Biemer, Lyberg, 2003, section 9.2]. Эти утверждения верны для парадигмы статистического вывода, основанного на проектировании (design-based paradigm). Однако из этого вовсе не следует, что для неслучайных выборок невозможны никакие статистические выводы. Поскольку случайные и неслучайные выборки используют различные процедуры для расчёта статистических характеристик и производства статистического вывода, в некоторых организациях оценки, получаемые по неслучайным выборкам, описываются в отличной от принятых для случайных выборок терминологии. Например, Национальный сервис статистического учёта в сельском хозяйстве (National Agricultural Statistics Service, NASS) [Matthews, 2008] называет оценки, генерируемые в неслучайных выборках, индикаторами (indications). Купер и Босняк [Couper, Bosnjak, 2010] определяют оценки в неслучайных выборках как меры внутренней валидности для выборки, но не для совокупности.

---

[15] Термины «изучаемая» и «целевая» совокупность являются синонимами. В настоящем переводе чаще используется термин «изучаемая», чтобы не создавать аллюзию на термин «целевая выборка», который описывает одну из разновидностей неслучайных выборок. Вместе с тем в маркетинговых исследованиях для обозначения изучаемой совокупности часто используется понятие «целевая группа», поэтому не следует полностью отказываться от термина «целевая совокупность». – *Прим. ред.*

Всякий раз, когда мы делаем выводы об изучаемой совокупности на основе выборки, статистические характеристики оценок становятся весьма значимыми. Особый интерес представляют смещение (оценки не смещены, если в среднем они равны параметрам совокупности); дисперсия (разброс оценок вокруг их среднего значения); среднеквадратическая ошибка (mean square error – показатель общей точности измерения, равный квадрату смещения плюс дисперсия). Поведение смещения и дисперсии с ростом размера выборки особенно важно, поскольку мы хотим, чтобы большие выборки были как можно более точными. Кратко остановимся на этих концептах.

Исследователи разработали разнообразные техники для улучшения статистических свойств выборочных оценок параметров совокупности. Для случайных выборок эти техники подробно описаны в учебных пособиях и в журнальных статьях по выборочным методам (см., например, [Sarndal, Swensson, Wretman, 1992]). Что касается неслучайных выборок, то результаты сильно различаются в разных дисциплинах – частично из-за большого разнообразия методов построения неслучайной выборки. Каждый из этих методов имеет свой набор литературы. Исследователи из одной области иногда заимствуют методы из других областей для решения проблем, связанных с получением статистических выводов из неслучайных выборок.

Некоторые из этих методов оценок или корректировок обсуждаются в других главах (выравнивание выборки, сетевая выборка). В этом разделе мы предлагаем несколько более общий взгляд на методологию оценки, включающую анализ весов и процедур, которые могут применяться для широкого круга неслучайных методов отбора. Мы опускаем бóльшую часть технических подробностей, но оставляем ссылки на работы – для тех, кто хочет детально разобраться в обсуждаемых вопросах.

6.1. СТАТИСТИЧЕСКИЕ СВОЙСТВА ОЦЕНОК

Выборки обеспечивают практический метод для оценивания параметров совокупности, но эти оценки зачастую не дают точных значений соответствущих

параметров. Источником этого несовершенства являются ошибки выборки (возникающие из-за того, что опрашивается не всё население) и ошибки, не связанные с выборкой (возникающие из-за неадекватных приёмов измерения). Даже при том, что эти ошибки в равной степени относятся к случайным и неслучайным выборкам, последние оцениваются более критично из-за общей непривычности неслучайных методов. (См. раздел 7, где подробнее обсуждается качество оценок для случайных и неслучайных выборок.)

Смещение – важнейший показатель качества всех массовых опросов, независимо от способа организации выборки. Смещение – это разница между усреднённой оценкой некоторого параметра и его значением в совокупности. Поскольку это определение приложимо как для случайных, так и для неслучайных методов отбора, разные методы получения статистических выводов требуют разных способов расчёта усреднённых оценок.

В случайных выборках статистические свойства оценок основаны на теоретическом среднем этих оценок, рассчитанном по всем возможным случайным выборкам, которые могут быть получены из заданной основы выборки по спроектированному дизайну[16]. Для расчёта теоретического среднего каждая оценка взвешивается в соответствии с вероятностью извлечения соответствующей ей выборки. В дизайне простой случайной выборки вероятность извлечения всех возможных выборок одинакова[17], поэтому усреднённая оценка считается без взвешивания как среднее арифметическое.

---

[16] Дизайн выборки определяется следующими параметрами: размером выборки, способом и вероятностями отбора респондентов из основы выборки, способом вычисления оценок (формулами для расчёта среднего, дисперсии, весовых коэффициентов и т. д.). В рамках одного дизайна можно получить огромное число случайных выборок, каждой из которых соответствует свой уникальный набор респондентов. Для каждой такой выборки можно рассчитать оценку того параметра совокупности, который интересует исследователя. Эти оценки могут различаться из-за разного состава респондентов в выборках. Усреднённая оценка (теоретическое среднее) получается путём вычисления среднего значения из оценок всех возможных выборок. Именно эта усреднённая оценка характеризует качество всего дизайна. Если она совпадает со значением параметра в совокупности, то дизайн выборки не имеет смещений (по данному параметру). При этом для конкретной реализации выборки оценка может не совпадать со значением в совокупности. Это обусловлено статистической погрешностью, присущей любой отдельной выборке, а величина возможного отклонения характеризуется шириной доверительного интервала. – *Прим. ред.*

[17] При размере совокупности N и размере выборки n существует C_N^n (число сочетаний из N по n) различных выборок. В дизайне простой случайной выборки каждую из них можно получить с одинаковой вероятностью, равной $1/C_N^n$. – *Прим. ред.*

Для стратифицированных и кластерных дизайнов расчёты более сложные, но и они хорошо описаны в соответствующих текстах, посвящённых теории выборки [Kish, 1965, section 1.3; Valliant, Royal, Dorfman, 2000, section 1.3]. Таким образом, смещение есть разница между усреднённой оценкой из всех возможных выборок и значением в изучаемой совокупности.

В неслучайных выборках отсутствие основы выборки и вероятностей отбора не позволяет рассчитать среднее по всем возможным выборкам в рамках разработанного дизайна. Исходя из тех же представлений, Департамент управления и бюджета США [OMB, 2006] использует для неслучайных выборок термин *ошибка оценивания (estimation error)* (см., например, [Levy, Lemeshow, 2008, section 2.4]). Схемы расчёта средних, применяемые для определения смещения, зависят от метода построения неслучайной выборки. Например, для статистических подходов, опирающихся на моделирование (model-based statistical approaches), в большинстве стандартных учебников средние рассчитываются на основе принимаемой модели распределения оцениваемых характеристик. Можно предположить, что некоторая характеристика – скажем, доход – имеет статистическое распределение, подобное нормальному, с некоторыми средним и дисперсией, которые затем и рассчитываются. Для этого вовсе не обязательно обосновывать применимость формальных статистических распределений – требуется только предположить, что наблюдения взяты из распределения с ограниченными средним и дисперсией. Если метод расчёта среднего значения определён, смещение рассчитывается как разница между оценкой среднего и значением оцениваемого параметра в изучаемой совокупности.

Кроме методов, основанных на проектировании и моделировании, применяется ещё один подход к производству статистических выводов, называемый *байесовский метод (Bayesian method)*. В приведённом выше примере для анализа электорального онлайн-опроса использовались байесовские методы. Американская ассоциация исследователей общественного мнения выпустила бюллетень по этому вопросу [AAPOR, 2012]. Байесовский подход сильно отличается от описанных выше методов, основанных на проектировании или моделировании. Одно из ключевых различий заключается в том, что оба предыдущих подхода отталкиваются от предположения, согласно которому

оцениваемая величина является константной; например, количество занятного взрослого населения есть величина фиксированная. В байесовской методологии эта величина рассматривается в качестве случайной, и статистический вывод заключается в использовании выборки с целью как можно лучше определить параметры её распределения. Чтобы решить эту задачу, приверженцы байесовской методологии начинают с рассмотрения исходного (начального) или априорного распределения численности занятого взрослого населения. Это априорное распределение зачастую может быть неопределённым (неинформативным) в том смысле, что мы ещё мало о нём знаем. Данные из выборочного обследования собираются, и первоначальное распределение модифицируется с использованием байесовской теории, чтобы получить апостериорное распределение, которое обобщает то, что стало известно из выборки. Например, апостериорное распределение общего количества занятых может быть аппроксимировано через нормальное распределение. Среднее и дисперсия этого апостериорного распределения дает информацию об общем количестве занятых. Область, где сконцентрирована основная часть апостериорного распределения (область с наибольшей плотностью распределения), называется *байесовским доверительным интервалом (credibility interval*[18]*)*.

Хотя байесовский подход сильно отличается от методов, основанных на проектировании и моделировании, своим представлением результата в терминах распределения, а не фиксированного числа, на практике мы находим множество сходств. Мы не обсуждаем байесовские методы в этом отчёте в основном по причине ограниченности объёма, отпущенного на представление неслучайных выборок. Мы предполагаем, что байесовские методы могут применяться для неслучайных выборок так же эффективно, как и в других методах, что потребует дальнейшей экспериментальной работы.

[18] Отличие байесовского доверительного интервала от привычного доверительного интервала состоит в следующем: для него утверждение, что истинное значение параметра совокупности попадает в вычисленный доверительный интервал с заданной вероятностью (95 % или какой-либо другой), справедливо только в том случае, когда для вычисления интервала использовалась адекватная модель распределения. Но в отличие от случайной выборки, тут не существует надёжного способа оценки, адекватна ли использованная модель, и если да, то в какой степени. – *Прим. ред.*

Задача построения выборки – получение оценок с небольшим уровнем смещений, где «небольшой» может пониматься по-разному в зависимости от целей исследования. В целом оценки со смещениями, которые уменьшаются с ростом размера выборки (состоятельные оценки), весьма желательны. На практике небольшие смещения могут определяться более утилитарно. Например, Олсон [Olson, 2006] определяет небольшое смещение как величину, которая статистически не отличается от нуля в анализе смещений из-за неответов (nonresponse bias analysis). Другие декларируют, что важны лишь существенные по величине смещения.

Смещение, как оно определено выше, – это не только результат применения выборки, в которую попадает лишь часть населения. Оно также включает невыборочные ошибки, такие как ошибки измерения. Оценки, которые усредняются в расчётах, являются фактическими оценками, а не некоторыми теоретическими значениями, связываемыми с совокупностью.

Неслучайные выборки имеют собственные источники ошибок, которых нет в случайных выборках (или по крайней мере не должно быть). Такие ошибки называют *ошибками отбора (selection bias)*. Например, в некоторых квотных выборках интервьюеров могут просить отбирать респондентов по указанным половозрастным группам, однако сам отбор остаётся за ними. Выбор интервьюерами конкретных респондентов, как правило, приводит к ошибкам отбора (см., например, [Lee, 2006]). Другие методы неслучайных выборок, такие как набор добровольцев, также приводят к ошибкам отбора [Bethlehem, 2010]. Важнейшее преимущество случайных выборок состоит в том, что они позволяют избежать этих смещений на этапе отбора (sampling stage), однако на этапе ответов (response stage) смещения могут возникать как в неслучайных, так и в случайных выборках.

Смещение — важнейшая характеристика выборочных оценок, поскольку оно может оказывать сильнейшее воздействие на их валидность. Например, Кокрен [Cochran, 1977] демонстрирует, как смещения могут приводить к такому доверительному интервалу, который включает истинное значение совокупности с гораздо меньшей вероятностью, чем заявленный уровень

доверия[19]. Более того, смещения из-за невыборочных ошибок зачастую не уменьшаются с увеличением размера выборки, и доверительные интервалы смещённых оценок при больших выборках имеют даже меньший уровень покрытия[20], чем при малых.

Дисперсия (variance) оценки – это среднеквадратическое отклонение оценки от её среднего (то есть среднее значение квадрата разности между оценкой и её усреднением). Дисперсия оценки тем ниже, чем меньше разброс между оценками, даже если ни одна из них не близка к оцениваемой величине из совокупности. Для случайных выборок при подходе, основанном на проектировании, дисперсия определяется как взвешенное среднее квадратов разностей между оценкой и её средним по всем возможным выборкам. В неслучайных выборках применяется тот же механизм усреднения, что и для оценки смещения.

Среднеквадратическая ошибка (mean square error) оценки – это скорее мера точности, а не разброса. Как таковая она выступает основной метрикой для сравнения разных выборочных методов и рассчитывается одинаково для случайных и неслучайных выборок; это квадрат смещения плюс дисперсия, но смещение и дисперсия оцениваются посредством разных механизмов усреднения для случайного и неслучайного дизайнов.

6.2. ПРОЦЕДУРЫ ОЦЕНКИ

Широкий набор процедур оценки, который применяется для неслучайных выборок, может быть разделён на две категории: основанные на (псевдо) проектировании (design-based) и основанные на моделировании (model-based). Обе категории представлены ниже.

---

[19] Например, вероятность того, что истинное значение совокупности лежит в заявленном 95%-ном доверительном интервале, может оказаться значительно меньше ожидаемых 95 %, скажем, всего 70 %. То есть вместо нужного уровня доверия 95% будет получен 70%-ный доверительный интервал. – *Прим. ред.*

[20] Здесь уровень покрытия означает то же, что и уровень доверия или доверительная вероятность. – *Прим. ред.*

Оценки, основанные на псевдопроектировании (pseudo design-based estimation). Статистической основой этого подхода является представление о вероятности попадания в выборку как единственном вероятностном компоненте оценивания в опросе. Этот процесс случайного отбора является основой для вычисления или усреднения по всем возможным выборкам смещения, дисперсии и среднеквадратической ошибки оценок. Статистический вывод об изучаемой совокупности опирается на процедуру случайного отбора, а не на какие-либо случайности в оцениваемых переменных.

Вес, основанный на псевдопроектировании (pseudo design-based weight), иногда применяется в неслучайных выборках; «псевдо» – потому что оценки строятся иначе, чем принято в традиционном подходе, основанном на проектировании. Вероятности отбора неизвестны и не определены (нет явной связи между выборкой и основой выборки, даже если основа существует). Вместо известных вероятностей отбора применяются оцениваемые вероятности попадания в неслучайную выборку. Идея опирается на нестрогое допущение, что каждое наблюдение в выборке репрезентирует другие, не отобранные (или отказавшиеся отвечать) единицы. Как только псевдовес сформирован, подсчитываются другие оценки, такие как оценки отношений (ratio estimators), при этом псевдовес выступает в качестве эквивалента известной вероятности отбора в традиционной случайной выборке.

Метод оценивания вероятности отбора варьируется в зависимости от ситуации, но общим является то, что вероятность отбора отдельных категорий респондентов рассчитывается как отношение размера выборки к численности этой категории в совокупности (как это делается, например, при постстратификации). Ясно, что создание или оценка псевдовесов требуют сильных предположений для некоторых неслучайных методов рекрутирования в выборку.

Оценки, основанные на моделировании (model-based estimation), опираются на статистическую модель, которая описывает оцениваемые в опросе переменные как имеющие нормальное распределение. В такой процедуре оценки предполагается, что интересующая характеристика (y-переменная) является случайной величиной, имеющей некоторое распределение, поэтому случайность возникает не в процессе создания неслучайной выборки.

Когда респонденты обследуются, то наблюдаемые данные считаются соответствующими модели, и последующий анализ проводится без учёта того, как была организована выборка. Другими словами, получаемые из модели оценки статистически не связаны с методом отбора. Это близко к тому, что можно найти в стандартных учебниках по статистике, где предполагается, что распределения в анализируемых данных аналогичны нормальному и соответственно можно оценить среднее и дисперсию распределения и посчитать другие статистические оценки. Применение такого подхода обычно требует независимого отбора наблюдений из всего интересующего исследователя распределения; и это предположение обычно нарушается, если, например, нас интересуют характеристики всех американцев, а опрашиваются жители одного штата.

Статистические модели часто не нуждаются в предположении о конкретном виде статистического распределения. Возможно, важнее, чтобы выполнялись условия по отношению к сопутствующим переменным, или ковариатам (например, одно из предположений заключается в том, что если значения сопутствующих переменных (ковариат) одни и те же, то наблюдения извлекаются из одного и того же распределения). Этот основанный на моделировании анализ и условия, при выполнении которых можно игнорировать метод отбора, но при этом делать валидные выводы, обсуждались рядом исследователей (см., например, [Little, Rubin, 2002; Smith, 1984; Pfefferman, Rao, 2009]).

Двухшаговая модель Хекмана [Heckman, 1976] – пример основанного на моделировании метода, взятый из эконометрической литературы. В этом подходе явно моделируются механизм отбора и итоговая переменная путём использования регрессионных моделей на каждом шаге. Первый шаг основан на наличии инструментальной переменной, позволяющей исправить какие-либо смещения отбора по искомой переменной. Инструментальная переменная должна быть тесно связана с анализируемой переменной, но не должна быть подвержена смещениям отбора, которые влияют на анализируемые переменные (см., например, [Fuller, 1987]). В отличие от большинства традиционных моделей измерения ошибок, в подходе Хекмана инструментальной переменной служит вероятность (предрасположенность) попадания в выборку, которая оценивается посредством первоначальной

регрессионной модели. Эта оцениваемая вероятность используется на втором шаге модели для корректировки смещений отбора (selection bias). Таким образом, регрессионные модели для этих двух шагов взаимосвязаны. Чтобы метод позволял производить эффективные оценки, необходимы некоторые строгие предположения (например, нормальность, корректное построение модели), незначительные отклонения от них могут приводить к нестабильным или смещённым оценкам.

6.3. ВЗВЕШИВАНИЕ

В случайных выборках взвешивание является неотъемлемой частью формулы оценки, построенной по набору ответов в опросе. Расчёт весов для случайных выборок начинается с расчёта базисных весов (base weights), иногда называемых проектируемыми весами (design weight) или весами, обратными вероятности отбора. Затем рассчитываются поправочные (корректирующие) веса, предназначенные для повышения точности оценок или для исправления возможных смещений, источником которых могут служить ошибки неответов или покрытия[21]. Калтон и Флорес-Сервантес [Kalton, Flores-Cervantes, 2003] описывают корректировки, направленные на уменьшение невыборочных ошибок (nonsampling error) в случайных выборках.

В неслучайных выборках веса могут применяться или не применяться, но их применение всегда должно опираться на практические соображения, и они не должны применяться только из-за того, что исследователь придерживается процедур, основанных на проектировании (design-based procedures). Понятно, что базисные веса, которые представляют обратные значения вероятностей отбора, не могут быть рассчитаны для неслучайных выборок, поскольку этих вероятностей не существует. Несмотря на это, для неслучайных выборок предлагается несколько подходов производства оценок, которые включают взвешивание в той или иной форме. Ниже мы обобщили эти подходы.

[21] Поправочные веса считаются с учётом значений базисных весов. Итоговый вес для каждого респондента получается путём перемножения его базисного и поправочного весов. – *Прим. ред.*

Без весов. В эконометрике и психометрии – двух областях, редко затрагиваемых в контексте массовых опросов, – для сбора данных применяются анкеты. В этих областях и в математической статистике обычно рассчитывают на случайные выборки (например, на простую случайную выборку, simple random sample – SRS) из целевой группы (см., например, [Casella, Berger, 2002]). Исходя из допущений простой случайной выборки для расчёта таких статистических данных, как средние, никакие веса не требуются. По существу, это форма псевдовесов, где все псевдовеса одинаковы и равны константе (в частности, единице).

В более сложных расчётах для снижения зависимости от строгих предположений простой случайной выборки применяются математические модели, которые содержат ключевые сопутствующие переменные (ковариаты). Применимость допущений простой случайной выборки для неслучайного отбора весьма проблематична, поскольку неясно, как определяются и отбираются потенциальные респонденты.

Корректировки по степени склонности (propensity score adjustments, PSA) – один из методов взвешивания, которым пытаются устранять смещения в случайных и неслучайных выборках. Как обсуждалось более детально в разделе 4, впервые корректировка по степени склонности (PSA) была представлена в исследованиях, построенных на наблюдениях (observational studies) [Rosenbaum, Rubin, 1983], как попытка сбалансировать известные параметры выборки для сопоставляемых групп (например, экспериментальной и контрольной), после того как измерения уже завершены.

Корректировки по степени склонности (PSA) (иногда представляемые как логистическое регрессионное взвешивание, logistic regression weighting) широко применяются в случайных выборках для уменьшения ошибок неответов. При этом исследователи исходят из предположения, что ответ представляет собой случайную величину [Little, 1986, 1988; Fricker, Tourangeau, 2010]. В этом случае модели ответа строятся с использованием логистической регрессии, где предсказывающие переменные (предикторы) известны как для респондентов, так и для нереспондентов. Полученная оценка склонности к ответам (то есть условная вероятность ответа) используется для

корректировки базисных весов респондентов в качестве коэффициента, на который умножается базисный вес и который, по предположению, является вероятностью участия респондента в опросе. Техническую информацию можно найти по ссылкам, приведённым выше, и в работе Бетлехема, Коббена и Шотена [Bethlehem, Cobben, Schouten, 2011, section 11.2.3].

Методы корректировки по степени склонности (PSA) также применяются в неслучайных выборках, особенно в опт-ин-панелях. Иногда они применяются для корректировки смешанных эффектов от ошибок покрытия, неответов и неслучайного отбора. В работе Ли [Lee, 2006] представлен расширенный список литературы об этом. Для оценки условной вероятности ответа в зависимости от всех этих источников может потребоваться проведение контрольного исследования по случайной выборке. Используя данные из обеих выборок, логистическая модель позволяет оценить вероятность участия в исследовании, построенном на неслучайной выборке. В идеале контрольный опрос должен строиться на случайной выборке, основа которой полностью покрывает изучаемую совокупность, где отсутствуют неответы и другие типы смещений [Bethlehem, Biffignandi, 2012, chapter 11]. Однако часто корректировка по степени склонности (PSA) опирается на относительно небольшие выборки, построенные методом случайной генерации стационарных телефонных номеров (RDD), с относительно низкими коэффициентами участия и проблемами покрытия, поскольку люди всё чаще переходят только на мобильные телефоны [Smith, 2011]. Потенциальные смещения контрольного исследования могут создать дополнительные проблемы при реализации корректировки по степени склонности (PSA) на неслучайных выборках (см. [Schonlau et al., 2003]).

Когда возможно проведение хорошего контрольного исследования для корректировки по степени склонности (PSA), фокус смещается на моделирование с применением вспомогательных (ковариаты) или предсказывающих (предикторы) переменных. Вспомогательные переменные (ковариаты) могут включать: демографические характеристики, которые выступают источником для многих постстратификационных корректировок; общие аттитюдные вопросы, известные как вебографические характеристики (webographic

characteristics[22]) [Duffy et al., 2005]; и опираться на наблюдения и данные, получаемые в ходе опроса, известные как параданные (paradata) [Groves, Lyberg, 2010].

Исследования по оценке эффективности вебографических вопросов (см., например, [Lee, 2006; Schonlau, van Soest, Kapteyn, 2007]) и параданных (см., например, [Kreuter et al., 2011]) для моделирования склонности были не очень успешными.

Валлиант и Девер [Valliant, Dever, 2011] изучили предположения, лежащие в основе корректировки по степени склонности (PSA) для неслучайных выборок. Они рекомендуют следующее:

- Вероятность быть включённым в неслучайную выборку и принять участие в опросе иногда может быть эффективно смоделирована с использованием переменных, заполненных в обоих исследованиях (по случайной и неслучайной выборкам).

- Корректировка по степени склонности (PSA) должна быть выполнена по модели, которая включает веса контрольного исследования. Первоначальные веса для элементов неслучайной выборки обычно приравниваются к единице, так что они репрезентируют исключительно респондентов, попавших в выборку. Однако некоторые исследователи полагают, что наиболее адекватный подход заключается в том, что перед запуском модели надо провести постстратификацию неслучайной выборки, вычислив веса, которые приведут её в соответствие со структурой совокупности (например, [Loosveldt, Sonck, 2008]).

- Если смещение в ответах на какой-либо вопрос связано с фактической вероятностью получения ответа (участия в опросе) и если

[22] Вебографики — это аттитюдные переменные, которые строятся для определения различий между людьми, участвующими и не участвующими в онлайн-опросах. Они обычно замеряют особенности стиля жизни, такие как увлечения (активности) людей и их частота за определённый период, потребление медиа, установки о приватности и открытость инновациям.

ответы на этот вопрос есть только для неслучайной выборки, то такое смещение не будет скорректировано с помощью этой техники. Даже если схожесть в ответах на вопросы для неслучайной и контрольной выборок будет высокой, ошибки отбора могут повлиять на некоторые ответы, полученные по неслучайной выборке. Это наблюдение может объяснить неодинаковые результаты относительно уровня смещений, обнаруженных на сегодняшний день (см., например, [Malhotra, Krosnick, 2007; Loosveldt, Sonck, 2008]).

Калибровочные корректировки (calibration adjustments). Весовая калибровка случайных выборок всесторонне изучается, и есть подтверждения, что она приводит к снижению как смещения, так и дисперсии в опросных оценках [Deville, Sarndal, 1992; Kott, 2006]. Два хорошо известных и широко применяемых примера калибровки – это постстратификация (poststratification) и выравнивание. Постстратификация весьма популярна как в случайных, так и в неслучайных выборках.

Другой метод калибровки, популярный в Европе и часто упоминаемый в американской литературе по опросам, известен как обобщённое регрессионное взвешивание (generalized regression weighting, GREG). Посредством линейной регрессионной модели конструируются калибровочные веса с использованием вектора вспомогательных переменных (auxiliary variables), которые известны для каждого элемента выборки и для совокупности в целом. Калибровочные веса рассчитываются таким способом, чтобы выборочные оценки с использованием этих весов были равны известным параметрам совокупности по каждой вспомогательной переменной. Выполняемая настройка описывается весом, который является функцией от коэффициентов линейной регрессии (если применяется линейная калибровка). Первоначальную версию подхода предложили Девилл и Сарндал [Deville, Sarndal, 1992] в качестве попытки скорректировать базисные веса в случайных выборках для увеличения точности измерений. Позднее подход был расширен и включил корректировку по неответам и уровню покрытия.

Калибровочные методы традиционно применяются к квотным выборкам; часто выборка взвешивается так, чтобы её значения совпадали с параметрами

совокупности по демографическим переменным, таким как национальность, возраст и пол. Наиболее часто применяется одна из разновидностей калибровки – постстратификация. Во многих неслучайных выборках постстратификация остается единственной формой взвешивания, в то время как для случайных выборок калибровка является дополнительным инструментом настройки весов, сформированных на основе базисных весов и, возможно, весов, выравнивающих неответы. Этот тип методологии взвешивания может оказаться единственным полезным инструментом для тех неслучайных выборок, у которых отсутствует необходимая информация для проведения корректировки по степени склонности (PSA) (выборка без соответствующего контрольного исследования) и у которых отсутствует контроль за отбором, подобный тому, что используется при сопоставлении несетевой и сетевой выборок.

Девер, Рафферти и Валлиант [Dever, Rafferty, Valliant, 2008] обнаружили некоторые, пусть и не вполне консистентные, достоинства обобщённого регрессионного взвешивания (GREG) для уменьшения смещения непокрытия (non-coverage bias). Перечисляя проблемы, связанные с корректировкой по степени склонности (PSA), Ягер с коллегами [Yeager et al., 2011] обнаружили, что в неслучайных выборках постстратификация увеличивает точность оценок, хотя результаты опять остаются нестабильными. В исследовании, проводимом Ли [Lee, 2006], и последующей работе Ли и Валлиант [Lee, Valliant, 2009] показано, что проводимые по отдельности корректировка по степени склонности (PSA) и калибровка, как правило, недостаточны для снижения смещений в оценках из неслучайных выборок до относительно низкого уровня.

Туранжо с коллегами [Tourangeau at al., 2013] обобщили результаты восьми исследований, в которых посредством взвешивания делались попытки снизить смещения, связанные с покрытием и эффектами отбора, в неслучайных опт-ин-панелях. Они обнаружили следующее.

- Корректировки позволяют избавиться лишь от части смещений, чаще всего – около 3/5.
- Корректировки иногда увеличивают смещения по сравнению с невзвешиваемыми оценками, порой более чем в два раза.

- Относительные смещения, которые остаются после корректировки, могут быть весьма существенными, часто сдвигая оценки на 20 % и более.
- Наблюдаются значительные различия между переменными, корректировки могут иногда устранять смещения, а иногда могут их значительно увеличивать.

В целом корректировки до некоторой степени снижают смещения, но не позволяют полностью избавиться от ошибок покрытия, неответов и отбора, присущих опт-ин-панелям.

Другие весовые корректировки. Эллиот [Elliott, 2009] разработал шкалированный «псевдовесовой» подход (scaled „pseudo-weights“ approach) для объединения неслучайной и случайной выборок. Его метод отличается от использования случайной выборки в качестве контрольной (reference sample) для корректировки по степени склонности (PSA), как описано выше. Его цель была объединить два опроса и анализировать их как одно исследование. Дальнейшее развитие методологии представлено в работе Эллиота и Хавиленда [Elliott, Haviland, 2007]. Они создали комбинированные оценки, объединяющие случайную и неслучайную выборки, где оценки из неслучайной выборки рассчитывались с помощью псевдовесов. Их метод оценивания близок байесовским методам, где оценки комбинируются таким образом, чтобы присвоить больший «вес» более точной оценке. Особый алгоритм, который они применяли, был разработан для техник статистический оценки малых групп (small area estimation; см. например, [Rao, 2003]). Они показали, что предлагаемый метод имеет меньшую среднеквадратическую ошибку, нежели аналогичный метод взвешивания, основанный лишь на контрольной выборке (reference sample).

6.4. ОЦЕНКА ДИСПЕРСИИ

Оценка дисперсии в случайных выборках опирается на подход, основанный на проектировании (design-based approach), в котором ключевую роль

играют вероятности отбора и расчёт среднего по всем возможным случайным выборкам с таким дизайном. В отношении неслучайных выборок опубликовано куда меньше работ, посвящённых оценке дисперсии, и основное внимание уделяется смещению. Однако для обоснования статистического вывода и оценки качества получаемых в опросах значений требуется куда больше исследований в такой важной области, как оценивание дисперсии. В результате отсутствия таких исследований утверждения, подобные заявлению Национальной сельскохозяйственной статистической службы (National Agricultural Statistical Service) [USDA, 2006] о невозможности производства валидных выводов, основанных на проектировании оценок разброса в неслучайных выборках, приводят к уверенности в том, что неслучайные выборки вовсе не позволяют рассчитывать дисперсию.

Весьма популярен подход, при котором предполагается, что была использована простая случайная выборка, и стандартное отклонение рассчитывается по формуле, применимой только для этого дизайна. Этот подход иногда применяется и для случайных выборок, но из литературы очевидно, что игнорирование конкретных выборочных дизайнов приводит к смещённым оценкам точности исследования. Аналогичный вывод, безусловно, относится и к неслучайным выборкам.

Поскольку оценки дисперсии, основанные на проектировании (design-based variance estimates), неприменимы для неслучайных выборок, для них была разработана техника расчёта дисперсии ошибки. Томпсон [Thompson, 1990, 2002] описал методы, основанные на псевдопроектировании (pseudo design-based methods), для специализированных адаптивных дизайнов выборки. Исакссон и Ли [Isaksson, Lee, 2005] предложили возможные техники, включающие постстратификацию по степени склонности (propensity score poststratification). Наконец, де Мунник, Дюпюи и Иллинг [Munnik, Dupuis, Illing, 2009] предложили технику повторной выборки (resampling technique).

Методы, основанные на псевдопроектировании (pseudo design-based methods). Томпсон [Thompson, 1990, 2002] описал модифицированные алгоритмы оценки Хансен–Гурвитца (Hansen-Hurwitz) и Хорвитца–Томпсона (Horvitz-Thompson), в которых учитывается вероятность включения в выборку

единиц, найденных через их связи с первоначальной случайной выборкой. Авторы предлагают использовать для оценки дисперсии формулу простой (стратифицированной) случайной выборки с модифицированными значениями. Отметим, что этот подход, основанный на простой случайной выборке, был доработан другими исследователями путём добавления оценочных весов, как обсуждалось выше.

Стратификация по степени склонности (propensity score stratification) – это техника, в которой оцениваемые вероятности отбора используются в постстратификации в дополнение к анализируемому весу. Способ оценки постстратифицированной дисперсии применяется в двух из трёх подходов, протестированных Исакссоном и Ли [Isaksson, Lee, 2005]. Первый и второй подходы используют методы постстратификации, включая модификацию метода, основанного на моделировании (model-based modification). В третьем подходе выборка случайным образом делится на группы равного размера, и затем для оценки дисперсии применяется стандартный метод «выкидного ножа» (jackknife[23]). Этот последний подход дает завышенную оценку истинной дисперсии.

Метод повторной выборки (resampling). Для оценки дисперсий широко применяются бутстреп-методы [Efron, Tibshirani, 1993]. Общий подход заключается в получении большого количества случайных подвыборок из первоначальной выборки, далее – в оценке интересующих параметров в каждой из подвыборок, и затем в определении дисперсии этих оценок с использованием методов аппроксимации Монте-Карло. Для случайных выборок Рао и Ву [Rao, Wu, 1988] предложили бутстреп-подход, который может применяться для выборочных опросов. Он применялся для неслучайных выборок в исследовании де Мунника, Дюпюи и Иллинга [de Munnik, Dupuis, Illing, 2009], но качество полученных результатов не анализировалось.

---

[23] Методом «выкидного ножа» (jackknife) из выборки опроса создаётся большое число подвыборок меньшего размера. Подвыборки получаются следующим образом: из каждой группы, на которые поделили выборку, удаляется по одному случайно отобранному элементу (элемент выбрасывается из выборки подобно лезвию выкидного ножа – отсюда и название метода). Для каждой полученной подвыборки считается своя оценка, а затем вычисляется дисперсия всех этих оценок, которая и используется в качестве итоговой оценки дисперсии. – *Прим. ред.*

6.5. ВЫВОДЫ

Основная озабоченность, связанная с неслучайными выборками, заключается в том, что оценки сильно зависят от допущений модели, на которой строится выборка. Эти допущения могут быть как простыми, так и сложными, и модели бывают скорее неявными, неоднозначными (имплицитными), чем точными и подробными (эксплицитными). Если допущения модели являются достаточно правдоподобными приближениями, то оценки, получаемые в неслучайных выборках, ведут себя хорошо, в том смысле что содержат ожидаемый (низкий) уровень смещения и дисперсии [Valliant, Dever, 2011].

Сложность заключается в том, чтобы понять, когда модели дают хорошие аппроксимации. Отсутствие связи (вероятности отбора) между выборкой и изучаемой совокупностью сильно затрудняет эту задачу. В случайных обследованиях эта связь используется для описания того, какие единицы совокупности недоступны для отбора (недопокрытие) и какие характеристики связаны с ошибками неответов среди тех, до кого удалось добраться через процедуры отбора.

Задача квантификации качества оценок в неслучайных выборках – архисложная. Сейчас применяются техники, решающие эту задачу для случайных выборочных обследований, и разработано несколько новых техник специально для неслучайных выборок. По мере апробации и распространения этих техник на практике ситуации, в которых применение неслучайных выборок наиболее оправданно, станут намного более понятными.

Пока этого не произошло, мы имеем широкий набор подходов, которые разрабатываются или уже применяются в областях, мало известных исследователям общественного мнения и маркетологам. С ростом осознания достоинств неслучайных выборок в отношении стоимости и быстроты проведения опроса мы ожидаем или как минимум надеемся, что ситуация изменится. В этой главе мы привели много таких подходов, однако затронули их весьма поверхностно.

7 Показатели качества опроса

Измерение качества данных, полученных с использованием неслучайных выборок, – новая непростая задача. В течение многих десятилетий исследователи измеряют качество данных способами, разработанными в парадигме случайной выборки. Многие опросные организации и государственные агентства разработали свои стандарты и руководства по проведению таких измерений. Обычно они сведены в специальный документ («профиль качества» – quality profile[24]), где описываются показатели качества и обобщается всё, что известно об источниках и параметрах ошибок. Поскольку эти показатели опираются на теорию вероятностей и применяются десятилетиями, они получили широкое признание как метрики качества.

К сожалению, неслучайные выборки нарушают три важнейших постулата, на которых базируются многие из этих показателей: 1) наличие основы выборки, содержащей все единицы изучаемой совокупности; 2) каждая единица имеет положительную вероятность попадания в выборку; 3) эта вероятность отбора может быть рассчитана для каждой единицы. Стандартные метрики качества спроектированы с целью измерять, насколько конкретная выборка отклоняется от этих предположений из-за ограничений, возникающих в ходе её реализации, таких как неполное покрытие и неответы.

Руководства по стандартам качества, разработанные Статистической службой Канады (Statistics Canada), Административно-бюджетным управлением США (U.S. Office of Management and Budget) и Бюро переписи населения

[24] «Профили качества» (quality profiles) в основном можно встретить у правительственных организаций. См., например: http://www.census.gov/sipp/workpapr/wp30.pdf.

США (U.S. Census Bureau), содержат «лишь краткие комментарии о методах оценки качества в неслучайных выборках. Например, Статистическая служба Канады допускает, что неслучайные выборки могут служить простым, быстрым и недорогим способом проведения предварительных исследований, фокус-групп и дополнительных опросов (follow-up surveys). Но далее говорится, что «возможность получения надёжных статистических выводов обо всём населении и оценки величины ошибки в этих выводах делает случайные выборки лучшим выбором для большинства статистических программ» [Statistics Canada, 2009].

Административно-бюджетное управление США опубликовало стандарты и руководства для федеральных статистических обследований [U.S. Office of Management and Budget, 2006]. В документе 20 стандартов, которые описывают все стадии исследования, включая разработку дизайна, пилотаж, сбор данных, их обработку и редактирование, анализ и распространение. Однако почти ничего не говорится о неслучайных выборках. Агентствам рекомендуется создавать выборки, опираясь на общепринятые статистические методы, например на «вероятностные методы, которые позволяют оценить ошибку выборки» [U.S. Office of Management and Budget, 2006, p. I]. Применение неслучайных выборок должно быть «статистически обоснованно и позволять измерить ошибки оценок»: «При применении методов построения неслучайных выборок следует включать их описание в документацию по дизайну исследования: какие опции рассматривались и обсуждались и почему был выбран именно этот дизайн, как проводилась оценка потенциальных смещений и какие методы использовались для их измерения. Дополнительно следует детально представить в документации процесс отбора и показать, что единицы, не попавшие в выборку, обоснованно исключены по объективным причинам» [U.S. Office of Management and Budget, 2006, p. 7].

Статистические стандарты качества, представленные в Бюро переписи населения США, тоже почти не касаются интересующего нас вопроса. В требовании A 3.3 читаем: «Основы выборок, отвечающие целям сбора данных, должны создаваться с применением *статистически обоснованных* методов». Логическое продолжение этого требования: «Статистически обоснованный дизайн выборки требует случайной выборки» [U.S. Census Bureau, 2011, p. 25].

Как мы уже не раз отмечали в этом отчёте, фактически во всех неслучайных методах отбора должны быть преодолены три препятствия: исключение большого количества людей из процесса отбора; опора на волонтёров или попавших в выборку по рекомендациям; и во многих случаях – высокий уровень неответов. При использовании одних методов, например конформных выборок, эти проблемы в основном игнорируются, при использовании других, таких как выравнивание выборки и разные виды постопросных корректировок, прикладываются значительные усилия для определения правильного набора вспомогательных переменных, которые помогут скорректировать выборку так, чтобы она репрезентировала (хотя бы приблизительно) изучаемую совокупность. Однако на момент написания этого доклада каких-либо общепринятых показателей или практик для проверки допустимости предположений и эффективности корректировок нет.

Отсутствие таких показателей и практик их применения, возможно, основная причина неохотного использования неслучайных методов отбора исследователями, особенно теми, кто работает на национальном уровне и задаёт стандарты опросной индустрии. Более того, даже в Американской ассоциации исследователей общественного мнения (AAPOR) недавно упоминалось о существенных затруднениях в оценке валидности предположений, на которые опирается репрезентативность опт-ин-панелей, и поэтому в случаях, когда требуется получение точных оценок, предлагалось обращаться к случайным выборкам [Baker et al., 2011, p. 758]. Из-за отсутствия общепризнанных основ для оценки качества данных в неслучайных выборках исследователи могут продолжать опираться на привычную концептуализацию общей ошибки опроса [Groves, 1989]. В этой теоретической рамке ошибки делятся на две большие категории: ошибки ненаблюдения (отсутствия наблюдения, error of non-observation) и ошибки наблюдения (error of observation). К первым относятся ошибки, влияющие на репрезентативность выборки, ко вторым – ошибки, влияющие на качество измерений, которое определяется тем, как исследование спроектировано и реализовано. Поскольку основная критика неслучайных выборок связана с их нерепрезентативностью, мы в первую очередь рассмотрим ошибки ненаблюдения.

7.1. ОШИБКИ НЕНАБЛЮДЕНИЯ

Ошибки ненаблюдения возникают в результате различий между изучаемой совокупностью, основой выборки и самой выборкой. Они обычно группируются в три категории: ошибки покрытия, ошибки выборки и ошибки неответов.

Ошибка покрытия измеряет, насколько хорошо основа выборки отражает (покрывает) изучаемую совокупность. В идеальных условиях каждый представитель изучаемой совокупности включен в основу выборки и поэтому имеет шанс быть отобранным. Гроувз с коллегами утверждал, что «без хорошо составленной основы выборки невозможно оценить ошибку покрытия» [Groves et al., 2009, p. 84]. Поскольку в неслучайных выборках единицы, как правило, не отбираются из хорошо составленной основы, задача оценить ошибку покрытия представляется нерешаемой.

Однако это суждение опирается на парадигму случайного отбора и не может быть напрямую перенесено на неслучайные выборки. В случайных выборках коэффициент покрытия чаще всего используется для измерения ошибки покрытия. Коэффициент покрытия – это отношение оцениваемой по выборке численности совокупности (которая рассчитывается исключительно через веса, обратные вероятностям отбора и, возможно, скорректированные по числу неответов) к известной численности совокупности[25]. В опросах домохозяйств это обычно отношения демографических показателей, где в знаменателе находятся данные последней переписи населения.

Для неслучайных выборок могут быть рассчитаны аналогичные коэффициенты, но они не будут опираться на веса, обратные вероятности попадания в выборку, поэтому расчёт должен быть иным. Можно сравнить относительные коэффициенты неслучайной выборки с характеристиками совокупности по данным переписи. Например, предположим, что целевая группа исследования – потребители инъекционных наркотиков в каком-то административном округе. И для поиска и рекрутинга респондентов в столь сложной

[25] Он может вычисляться как для всей совокупности, так и для отдельных групп. – *Прим. ред.*

для опроса совокупности применяется выборка, управляемая респондентами (respondent-driven sample). Если клинические записи в этом округе корректно отражают медицинские характеристики населения, то демографические коэффициенты, такие как распределения по полу, возрасту, расе, национальности, могут быть рассчитаны. Если демографические коэффициенты выборки значительно различаются с демографическими коэффициентами, посчитанными по медицинским записям, это может означать наличие ошибки покрытия. Если же они совпадают, это не значит, что покрытие полное: можно лишь утверждать, что покрытие одинаково среди групп, различающихся по полу, возрасту, расе и национальности.

Основная проблема заключается в том, что когда по неслучайной выборке нужно оценить суммарные (абсолютные, totals) характеристики совокупности, то для получения эквивалента выборочных весов приходится полагаться на внешние данные о совокупности, поскольку отбор из основы выборки не проводится. Это приводит к значительным сложностям в расчёте абсолютных, а не относительных коэффициентов покрытия. Исследователи применяли разные подходы для расчёта абсолютных коэффициентов покрытия, но с переменным успехом. Например, один из таких подходов предполагает в качестве показателя покрытия в онлайн-исследованиях, использующих выборку из опт-ин-панели, рассматривать долю населения, имеющего доступ в интернет. Хотя этот показатель вполне информативен, он опирается на важное допущение, что все интернет-пользователи могут получить приглашение на участие в опросе и, кроме того, подходят для участия в нём. Для очень многих ситуаций это просто неверно. Поэтому для неслучайных выборок необходимы более достоверные показатели покрытия.

Конечно, коэффициент покрытия, как и коэффициент ответов, – несовершенный показатель для смещений из-за неполного покрытия даже для случайных выборок. В качестве наиболее показательного примера предположим, что основа выборки – это список адресов в почтовой службе США, который покрывает 98 % населения страны и используется для построения случайной выборки. Коэффициент покрытия очень высок, но эта выборка будет давать никудышное покрытие для бездомных.

В недавней статье Блэра и Конрада [Blair, Conrad, 2011] обсуждается «проблема» покрытия в контексте пилотажных когнитивных интервью и акцентируется внимание на определении размера неслучайной выборки для адекватного понимания проблем, связанных с анкетой. Для случайных выборок определение размера непосредственно связано с ожидаемым уровнем ошибок выборки; в отношении неслучайных выборок ситуация куда менее понятна. Блэр и Конрад предпринимают первую попытку расчета и составления практических рекомендаций по определению размера выборки. Они показывают, что различные объёмы квотной выборки (начиная с пяти и заканчивая 90 интервью) нужны для изучения разных характеристик (например, среднего количества зафиксированных проблем, вероятности обнаружения новых значимых проблем). Один из их выводов состоит в том, что «в малых выборках может быть пропущен значимый процент проблем, даже если ограничиться исследованием только тех проблем, которые существенно влияют на ошибку измерения» [Blair, Conrad, 2011, p. 654]. Необходимо проведение дополнительных исследований, подобных этому, чтобы адекватно оценить нужное количество единиц наблюдения для квотных и других неслучайных выборок.

Ошибка выборки. Следующий компонент модели общей ошибки опроса – ошибка выборки – измеряется дисперсией оценок, которые могут быть получены во всех возможных выборках, проводимых по одинаковому дизайну и использующих одну и ту же основу. Когда основа выборки обеспечивает полное или почти полное покрытие всей совокупности, резонно заключить, что такой способ измерения позволяет оценить (по крайней мере приблизительно) точность полученных результатов исследования. Однако предположение о полном покрытии редко применимо к неслучайным выборкам. Это особенно очевидно в отношении неслучайных панелей (non-probability panels), поэтому Американская ассоциация исследователей общественного мнения утверждает, что «указание интервалов ошибки выборки для опт-ин- или самообразующихся панелей заведомо ложно» [Baker et al., 2011, p. 773].

Мы согласны, что интервалы ошибки выборки несут определённые смыслы, и эти измерения неадекватны для неслучайных выборок. Однако в статистической литературе используются термины, которые прямо сопоставимы

с ошибкой выборки как способом измерения разброса в оценках и при этом никак не связаны с идеей всех возможных выборок. Например, в самом элементарном статистическом тексте предлагается модель (скажем, случайная выборка из нормального распределения) и рассчитывается стандартная погрешность; при этом модель рассматривается как мера неопределённости из-за применения выборки.

Мы уверены, что следует стимулировать разработчиков неслучайных выборок включать в отчёты информацию о точности их оценок, но во избежание недоразумений мы рекомендуем придерживаться терминологии, отличной от разработанной для случайных выборок. Точность оценок для неслучайных выборок – это не среднее отклонение по всем возможным выборкам, а скорее рассчитанное по модели отклонение от значения по населению в целом. Например, крупнейшее агентство маркетинговых исследований Ipsos предложило использовать байесовский доверительный интервал (credibility interval) [Ipsos, 2012] для оценки исследований, основанных на опт-ин-панелях. Как указывалось в разделе 6 настоящего отчёта, байесовский доверительный интервал – это мера неопределённости, используемая в байесовских методах, и Ipsos описывает свои процедуры как байесовские. Другие подходы, основанные на моделировании, также производят оценки точности, такие как стандартные ошибки, которые не имеют отношения к среднему по всем возможным выборкам (терминология, принятая для статистических выводов, основанных на проектировании, которая применяется в случайных выборках).

Хотя перечисленные методы не являются общепризнанными в исследовательском сообществе, мы уверены, что опросная индустрия нуждается в конструировании и развитии таких инструментов, которые позволят заполнить пробел, образовавшийся в результате непригодности стандартного способа расчёта погрешности для неслучайных выборок. Рассмотрение оценок, как если бы в них вовсе отсутствовали ошибки, не может быть признано допустимым решением. С этой точки зрения исследователи должны сами принимать решение о полезности тех или иных показателей. Оценка качества показателей возможна только тогда, когда исследовательские организации в соответствии с Кодексом профессиональной этики и практики Американской

ассоциации исследователей общественного мнения полностью раскрывают всю информацию и предоставляют детальное описание особенностей построенной модели, процедур валидации и методов расчёта показателей.

Конечно, существует много методов отбора для неслучайных выборок, и нет простого и универсального способа оценки ошибки. Например, Садмен [Sudman, 1966] предложил технику квотной выборки и показал, что если выборка осуществляется в соответствии с определёнными предположениями исходя из предложенных им процедур, то ошибки этой квотной выборки могут рассчитываться по тем же правилам, что и для случайной. Другие методы оценки ошибок выборки разработаны для отбора, управляемого респондентами (respondent-driven sampling). Вновь отметим: поскольку методы построения неслучайных выборок сильно разнятся, чрезвычайно важно предельно ясно документировать все нюансы процедур и предположений, необходимых для расчёта разброса или точности оценок.

Важно, чтобы методы построения неслучайных выборок оценивались исходя из разработанных для них показателей точности. Хорошо зарекомендовавший себя подход состоит в использовании повторных выборок (репликаций) – идея, возникшая на заре изучения случайного отбора [Mahalanobis, 1946; Deming, 1944]. Данный подход также применялся в наблюдениях и экспериментах. В простейшем виде он состоит в том, что создается несколько выборок (многократная выборка), и разброс (дисперсия) оценок между ними используется для измерения точности оценки. Эта идея получила развитие в методе псевдорепликаций (псевдоповторных выборок, псевдотиражирования выборок, *pseudo-replication*), когда исследование проводится по одной выборке, которая потом делится (тиражируется) на несколько подвыборок – репликаций. Теория псевдорепликаций для случайных выборок тщательно разработана [Wolter, 2007], и способы её приложения для неслучайных выборок вполне реализуемы. Репликации могут быть экономичным, относительно надёжным и понятным инструментом для многих практиков: они не требуют больших выборок и могут быть выполнены относительно быстро. Однако этот подход, как и традиционное вычисление погрешности для случайных выборок, позволяет лишь оценить разброс (дисперсию), а не измерить отклонения.

Ошибка неответа. Коэффициент ответов в случайной выборке – это отношение подходящих по условиям отбора единиц, принявших участие в измерении, ко всем подходящим по условиям отбора единицам, включённым в выборку [AAPOR, 2009]. Коэффициент ответов – вероятно, наиболее известный показатель качества для случайных выборок, хотя в последние годы его ограничения в качестве предиктора для смещений, связанных с неответами, стали более очевидными [Groves, 2006].

В неслучайных выборках знаменатель для расчёта коэффициента может быть неизвестен, поэтому не всегда можно рассчитать коэффициент ответов, как того требуют стандарты Американской ассоциации исследователей общественного мнения или других профессиональных сообществ. Соответственно, как и в случае с расчётом погрешности, исследователи, описывающие неслучайные выборки, должны избегать термина «коэффициент ответов» и вместо него использовать другие. Отраслевой стандарт ИСО 20525:2008 рекомендует термин «коэффициент участия» (participation rate), который определяется как «количество респондентов, которые ответили на вопросы, делённое на общее количество персональных приглашений принять участие в опросе» [ISO, 2008]. Этот термин был заимствован Американской ассоциацией исследователей общественного мнения и включён в издание Стандартных определений 2011 года [Standard Definitions, 2011]. Айзенбах [Eysenbach, 2004] предлагает набор других связанных показателей, включая коэффициент рассмотрения (view rate) и коэффициент заполнения (completion rate).

В опт-ин-панелях на этапе рекрутинга зачастую собираются детальные сведения об участниках, поэтому исследователи могут использовать эту информацию для оценки различий между респондентами и нереспондентами в отдельных исследованиях. Эти данные могут применяться для оценки потенциальных смещений, связанных с неответами, и анализа качества данных, но до последнего времени неизвестны примеры таких процедур.

Другие метрики ответов, отражающие специфику опт-ин-панелей, детально разобраны Коллегаро и ДиСогра [Callegaro, DiSogra, 2008]. Некоторые из них, но не все, могут быть применены для неслучайных панелей.

- *Коэффициент абсорбции (поглощения, absorption rate)* – доля электронных писем (e-mail), не доставленных адресатам из-за ошибочного адреса, переполненного почтового ящика или ошибок, связанных с сетью. Это индикатор того, насколько хорошо провайдер панели обновляет данные и взаимодействует с её участниками.

- *Коэффициент прерываний (break off rate)* – количество обрывов в заполнении анкеты[26]. Высокий уровень коэффициента прерываний может служить индикатором качества, сигнализирующим о проблемах с дизайном анкеты (плохое форматирование или навигация, слишком затянутая анкета).

- *Коэффициент завершения скрининга (screening completion rate) / коэффициент соответствия условиям отбора (study specific eligibility rate)* – количество людей, которые полностью прошли скрининг и подошли по условиям отбора, плюс количество людей, которые полностью прошли скрининг и не подошли по условиям отбора, делённое на общее количество приглашений к участию в опросе. Значительное отклонение коэффициента скрининга (соответствия) от устоявшихся значений в подобных исследованиях служит индикатором сфабрикованных / некорректных ответов участников панели, прошедших самоотбор для этого опроса.

- *Коэффициент износа (attrition rate)* измеряет процент участников панели, которые отказались от участия в течение установленного периода времени. Это расчёт числа участников панели, остающихся активными месяц за месяцем [Clinton, 2001]. Высокий коэффициент износа может сигнализировать о плохом качестве дизайна анкет (опрос слишком долгий) и в результате большей утомляемости и высоких коэффициентах отказов от дальнейшего участия в панели (dropout rates).

---

[26] В отчёте не указан знаменатель коэффициента. По всей видимости, авторы забыли его добавить. Можно предположить, что в качестве такового целесообразно считать общее количество анкет, которые начали заполнять участники опроса (включая полные и прерванные). — *Прим. перев.*

Относительно немного показателей было разработано для оценки онлайн-выборок. Большинство их них применимы для опт-ин-панелей – доминирующего инструмента в онлайновых маркетинговых исследованиях последнего десятилетия. Но опт-ин-панели быстро устаревают по мере роста спроса на онлайн-респондентов: клиенты ищут большее демографическое разнообразие в онлайн-выборках, увеличивается интерес к более узким группам населения. Провайдеры онлайн-выборок всё чаще опираются на различные источники, выходящие за рамки их панелей. Они обращаются к панелям конкурентов, социальным сетям, размещают приглашения участвовать в опросе на различных сайтах, выборка становится всё больше похожа на потоковую (river sampling). Респондентов теперь не приглашают принять участие в конкретном опросе на узко заданную тему – вместо этого они получают приглашение общего характера на участие в опросах вообще. Подтвердив своё согласие, участники перенаправляются на сайт, где отбираются и ставятся в очередь на тот опрос, для которого подходят. Программное обеспечение, которое контролирует этот процесс, называется роутером (router). Его задача – обеспечить каждого желающего подходящей анкетой. На сегодня существует уже множество таких программ, по-разному работающих и оказывающих разное воздействие (если оно имеется) на данные. К сожалению, роутеры также не позволяют рассчитывать коэффициент участия, о чём говорилось выше.

Изучение ошибок неответов вошло в большинство финансируемых государством исследований, основанных на случайных выборках. Административно-бюджетное управление США играет ключевую роль в развитии подходов к анализу смещений, связанных с неответами. Использование подвыборок нереспондентов может быть подходящим инструментом для некоторых видов исследований, основанных на неслучайных выборках. Техники, которые задействуют индивидуальные / географические внутренние характеристики, имеющиеся в основе выборки, широко используются в случайных выборках и могут быть применимы для неслучайных выборок при сборе информации о потенциальных респондентах. В некоторых случаях доступны надёжные внешние проверочные значения (benchmarks) по одной или нескольким ключевым дополнительным переменным, включённым в текущее исследование. Когда доступно немного внутренней или

внешней информации, появляется возможность оценивать ошибки неответов, используя методы внешней валидности. Например, в медицинской литературе говорится, что среди людей, отвечающих определённому условию, примерно половина – женщины. Если при неслучайном отборе таких людей (например, с использованием метода неслучайных стартовых точек – seeds) в выборке оказывается 80 % женщин, очевидно, что в процессе реализации выборки мужчин реже приглашали или они реже соглашались участвовать в опросе. В этой ситуации неясно, поможет ли взвешивание выборки по полу и другим социально-демографическим характеристикам добавить в нее выпавшие 30 % мужчин.

Аналитически подкрепляя сложность неответов в неслучайных выборках, Гайл, Джонстон и Салганик [Gile, Johnston, Salganik, 2012] представляют трёхуровневый показатель неответов для выборки, управляемой респондентами (respondent-drviven sampling). Основываясь на ответах респондентов о количестве купонов, по которым отказались отвечать их получатели, и общем количестве купонов, переданным другим, авторы рассчитали коэффициент купонных отказов (coupon-refusal rate) и коэффициент купонных невозвратов (coupon non-return rate), которые вместе образуют общий коэффициент неответов.

7.2. ОШИБКИ ИЗМЕРЕНИЯ

Другая большая категория ошибок, находящаяся в рамках модели общей ошибки опроса (TSE), связана с наблюдениями и чаще определяется как ошибки измерения. Обычно выделяют четыре источника этих ошибок: анкета, интервьюер (если он есть), респондент, метод опроса (личное интервью, почтовый опрос, телефонный или онлайн). Здесь мы можем согласиться, что данные, получаемые при использовании неслучайных выборок, похоже, имеют такие же ошибки, что и данные из случайных выборок. Оба типа имеют тенденцию накапливать разрывы наблюдения: между конструктами, которые необходимо замерить, и инструментарием, предназначенным для измерения; между использованным инструментарием и полученными

ответами; между полученными ответами и данными, записанными и отредактированными. Неважно, с помощью случайной или неслучайной выборки собираются данные. Мы должны разработать для них общие индикаторы, позволяющие фиксировать ошибки измерения.

Анкета. Один из наиболее распространённых индикаторов качества, связанный с анкетой массового опроса, – *конструктная валидность* (construct validity). Это степень соответствия вопросов или блоков вопросов анкеты измеряемым свойствам ключевых конструктов, или степень, с которой обследование измеряет истинное значение конструкта. Так, если мы хотим оценить отношение к конфиденциальности данных или информации, то будем ждать, что ответы респондентов, сказавших, что они «очень обеспокоены» (very concerned) конфиденциальностью данных, могут также включать обеспокоенность компьютерным воровством паролей, хакерами или распространением этих данных правительственными агентствами в рамках их рутинной деятельности. Если мы обнаружим, что эти вопросы коррелируют ожидаемым образом, мы получим некоторые гарантии корректного измерения конструкта. Мы также можем сравнить ответы на ключевые вопросы в разных группах, где ожидаются различия в ответах. Например, изучая установки в отношении таких проблем, как однополые браки, контроль оружия, ограничения на аборты, мы можем ожидать значимых различий между республиканцами и демократами.

Более прямой способ измерения конструктной валидности – повторное задавание ключевых вопросов в одной анкете, специально для этого спроектированной. В Национальном обследовании домохозяйств о потреблении наркотиков (National Household Survey on Drug Use) эта техника применяется для валидации ответов о потреблении марихуаны [Biemer, Lyberg, 2003]. Первым задаётся вопрос: «Когда вы последний раз употребляли марихуану или гашиш?» Позднее в этом же опросе респондента спрашивают: «Сколько дней за последние 12 месяцев вы потребляли марихуану или гашиш?» Если он отвечает на второй вопрос: «Один день или больше», — мы можем ожидать ответ на предыдущий вопрос: «Менее года назад». Высокий уровень несоответствия между ответами на эти вопросы может говорить о наличии проблем с измерением потребления марихуаны.

Интервьюер. Во многих современных исследованиях, опирающихся на неслучайные выборки, интервьюер вообще отсутствует, поскольку всё чаще используется самоадминистрирование – когда респондент самостоятельно заполняет анкету на бумаге или онлайн. Однако рассматривать интервьюера как источник ошибок измерения вполне уместно, когда опрос по неслучайной выборке ведётся лично или по телефону. Уже накоплен огромный объём литературы об эффекте интервьюера как источнике ошибок измерения для случайных выборок [Groves, 1989], и большинство публикаций (если не все) релевантны для исследований, проводимых на неслучайных выборках. В них описываются техники встроенных в опросы методических дизайнов для измерения различий, связанных с интервьюерами; анализируется связь между характеристиками интервьюеров и ответами на вопросы анкеты; оцениваются преимущества рандомизации интервьюеров и работы в условиях специально оборудованных централизованных звонковых центров; измеряется соблюдение интервьюерами инструкций и анализируются параданные для определения возможных случаев фальсификации данных.

Респондент. Все неслучайные выборки в опросах опираются на некоторую совокупность респондентов, собираемую неслучайным образом, и эти респонденты становятся потенциальными источниками ошибок измерения. Кан и Каннел [Kahn, Cannell, 1957] и Гроувз [Groves, 1989] определили пять стадий производства ответа.

1. Кодирование информации.
2. Понимание.
3. Воспоминание.
4. Поиск приемлемого ответа.
5. Коммуникация.

Фактически все методы оценки связанных с поведением респондентов ошибок, которые разработаны для случайных выборок, релевантны и для неслучайных. Обратная проверка записей (reverse record check) может использоваться для сравнения ответов респондентов с административными / программными / медицинскими записями. Этот метод также потенциально весьма эффективен на этапе разработки инструментария для определения продолжительности срока доступных респонденту воспоминаний (recall period). Когнитивное интервью может быть весьма полезно при оценке

ошибок измерения для большинства из перечисленных выше шагов. С его помощью можно успешно тестировать восприятие скрининговых критериев, особенно если для респондентов, участвующих в когнитивном интервью, из внешних источников известны характеристики, позволяющие определить их принадлежность к целевой группе. Это даёт возможность дополнительно протестировать скрининговые вопросы.

Широкое распространение интернета как области сбора данных позволяет получать детальную аналитическую информацию о взаимодействии респондента с анкетой. Помимо сведений о местах прерывания опроса хорошо спроектированные веб-анкеты могут также измерять время, затраченное на каждый вопрос, прямое и обратное перемещение по анкете, частоту выбора ответа «не знаю» как показателя доли неответов и длины ответов (например, количество слов) в открытых вопросах. Наконец, имеется целый набор приёмов оценки качества, разработанный для парадигмы опт-ин-панелей, который подробно освещается в п. 7.4.

Метод сбора данных. Среди факторов, оказывающих наиболее сильное влияние на эффект ответа, Гроувз [Groves, 1989] выделяет метод сбора данных. Сегодня применяются опросы, проводимые персонально; личные интервью с предоставлением респонденту коммуникативного электронного устройства для связи с централизованным опросным центром; почтовые опросы, децентрализованные и централизованные телефонные опросы, интернет-опросы. Картина становится намного сложнее, когда мы рассматриваем гибридные способы, такие как личные интервью, в которых респондент самостоятельно заполняет анкету на электронном оборудовании (компьютер, планшет, смартфон).

За последние 10 лет в разных исследованиях, основанных на случайных выборках, применялось большое количество разнообразных методов сбора данных. Рост различных мультимодальных исследований позволил проводить эксперименты с методами сбора данных, используемыми для случайных выборок. Основная цель таких экспериментов – определить, насколько каждый метод сбора данных приводит к эффекту метода. Литература по эффектам метода почтовых и веб-исследований наиболее релевантна

для неслучайных выборок, поскольку представляет два основных метода сбора данных. Некоторые типы неслучайных выборок пригодны для рандомизации выбора метода между почтовым и интернет-опросом или для предоставления этого выбора респонденту, который имеет возможность выбирать между двумя или более опциями. Также важно оценивать коэффициент прерванных интервью для мультимодальных исследований и различать прерывания во время скрининга и основной части анкеты.

7.3. ВНЕШНЯЯ ВАЛИДНОСТЬ

Идея внешней валидности – центральная в большинстве работ по оценке качества собранных в опт-ин-панелях данных. В публикациях последнего десятилетия описываются попытки оценить валидность выборок посредством сопоставления с внешними источниками: со случайными выборками (чаще всего телефонными), эталонными (benchmark) данными, такими как перепись населения, результаты выборов, и другими данными, собранными неопросными методами (например, административными или данными об объёмах продаж). Обзор этой литературы представлен в Отчёте рабочей группы AAPOR по опт-ин-панелям [Baker et al., 2011] и здесь не воспроизводится. Наверное, самая цитируемая из этого списка – работа Йегера и его коллег [Yeger et al., 2011]: они рассмотрели пять опт-ин-панелей и сопоставили их результаты по небольшому набору переменных с двумя случайными выборками: с высококачественным государственным обследованием и с административными записями.

На первый взгляд, это весьма эффективный способ оценки валидности неслучайных выборок. На практике такая оценка становится проблематичной как минимум по трём причинам.

Во-первых, это не всегда возможно, поскольку необходимые для сравнения внешние показатели могут отсутствовать. Для оценки качества можно проводить тщательно спланированные эксперименты, но они никогда не станут стандартными приёмами в подобных исследованиях. Если целевая

совокупность и тематика опроса регулярно находятся в центре исследовательского внимания и как-то отслеживаются, то вполне возможно найти готовые источники для сопоставления. Но как только целевая совокупность становится более узкой, а темы более закрытыми, найти такие источники становится крайне сложно, а то и вообще возможно.

Во-вторых, даже в высококачественных исследованиях, проведённых по случайным выборкам и замеряющих одни и те же параметры, результаты иногда расходятся. Например, количество незастрахованных людей в США замеряется как минимум в четырёх федеральных государственных обследованиях, каждое из которых имеет высокий коэффициент ответов. Тем не менее в оценках присутствуют различия, выходящие за пределы ошибки выборки, и, видимо, объясняются ошибками измерения или другими невыборочными ошибками [Davern et al., 2011]. Другой пример. В Обследовании доходов и участия в государственных программах (Survey of Income and Program Participation, SIPP) количество людей, прошедших тест по общеобразовательной подготовке (General Education Development test, GED), примерно на 70 % больше, чем по оценкам, полученным в Текущем обследовании населения (Current Population Survey, CPS) [Crissey, Bauman, 2012]. Оба источника – весьма качественные обследования, проводимые Бюро переписи населения.

Наконец, административные записи и другие данные, собираемые вне опросных методов, также небезошибочны. Например, Смит с коллегами [Smith et al., 2010] описывают ситуацию, когда примерно в 25 % случаев в административных записях имеются ошибки в классификации этничности и расы, в основном из-за пропусков информации. И это вовсе не исключительная ситуация в административных данных, которые не предназначены для статистического применения. Брэкстоун [Brackstone, 1987] обобщил достоинства и недостатки использования административных данных для этих целей. Борух [Boruch, 2012] также провел оценку различных типов административных записей и подтолкнул к весьма продуктивной дискуссии.

Подводя черту, можно отметить, что во многих случаях трудно проинтерпретировать различия в данных разных исследований и выделить из них ту часть расхождений, которая определяется особенностями

построения выборки. «Золотого стандарта» для таких сравнений пока нет. Тем не менее мы можем судить о валидности, полагая, что нам кое-что известно о качестве подлежащих сравнению данных или источников, из которых они получены. Но точная оценка ошибок, связанных с механизмами отбора, – задача чрезвычайно трудная.

7.4. ПОКАЗАТЕЛИ КАЧЕСТВА ДЛЯ ОПТ-ИН-ПАНЕЛЕЙ

Уже более 10 лет в индустрии маркетинговых исследований активно используются различные неслучайные методы, опирающиеся на онлайн-панели. Некоторые из этих методов описаны в предыдущих главах, и результаты исследований, проведённых с использованием опт-ин-панелей, часто проблематизировались на основании возникающих смещений и разброса [Yeager et al., 2011; Walker, Petit, 2009]. Однако как упомянутые, так и другие исследования обычно не затрагивают вопросы методов организации выборки. Основное внимание уделяется скорее панели как таковой, нежели техникам, применяемым для реализации выборок на её основе. С этой точки зрения представляется важным сфокусироваться на вопросах покрытия и различных способах, посредством которых панели рекрутируются и поддерживаются. Аналогичным образом расчёт ошибок покрытия и оценка техник, которые могут для этого применяться, – важнейшая часть анализа качества выборок, реализуемых в опт-ин-панелях.

Вместе с тем на сегодня общепризнано, что в индустрии маркетинговых исследований есть три дополнительных проблемы, возможно, уникальных и свойственных исключительно панельным моделям. Перечислим их.

- Люди склонны записываться в одну и ту же панель под разными идентификаторами, чтобы увеличить шансы быть отобранными для исследования.
- Когда исследование проводится с использованием выборок из разных панелей, некоторые люди могут быть отобраны и пройти один и тот же опрос больше одного раза. Более грубое

нарушение методики отбора заключается в использовании веб-роботов (в просторечии «ботов») для многократного автоматического заполнения анкеты.

- Частое применение стратегии минимизации усилий (satisficing)[27], признаками которой являются очень короткое время заполнения анкет, линейное прохождение табличных вопросов, частый выбор неподходящих ответов, пропуск вопросов, противоречивые или бессмысленные ответы.

Значимость этих проблем в отрасли и степень их влияния на результаты исследований являются предметом дискуссий, один из аспектов которых – утверждение о необходимости прибегать не только к разным панелям, но и к разным типам исследований. Кроме того, теперь пришло понимание важности проблем качества данных — этот вопрос поднимается опросными организациями, профессиональными ассоциациями, разработчиками программного обеспечения, компаниями, поддерживающими панели, и индивидуальными исследователями. Например, оба стандарта ИСО по исследованиям в целом и исследованиям, опирающимся на панели (ISO 20252 – Market, Opinion and Social Research; ISO 26362 – Access Panels in Market, Opinion and Social Research), определяют требования, которые должны быть удовлетворены и по которым должна производиться соответствующая отчётность. Сегодня рынок имеет два решения в области программного обеспечения для определения потенциально проблемных респондентов в онлайн-выборках: релевантный идентификатор (Relevant ID) и истинная выборка (true sample). Многие индивидуальные исследователи проводят процедуры редактирования данных после завершения полевых работ, чтобы исключить продублированных респондентов и потенциальных минимизаторов усилий (satisficers) [LeGuin, Baker, 2007].

Но эти процедуры поддержания качества выборок сами приводят к определённым проблемам. Например, процедуры валидизации

[27] Термин satisficing является комбинацией слов satisfy (удовлетворять) и suffice (быть достаточным). Им обозначают респондентов, которые отвечают на вопросы автоматически, не задумываясь над их содержанием, лишь бы побыстрее и с минимальными усилиями заполнить анкету. – *Прим. ред.*

идентичности предполагаемых панелистов, как показывает практика, исключают из панели людей младше 30 лет, небогатых, менее образованных, небелых гораздо чаще, чем представителей других демографических групп [Courtright, Miller, 2011]. Также нет чётких стандартов для измерения степени соответствия респондента условиям отбора (включения в исследование) или для того, чтобы отличить плохое исполнение роли респондента от плохого дизайна анкеты.

Вместе с тем исследователи, заинтересованные в качественных выборках из опт-ин-панелей, благоразумно требуют официальных отчётов, подтверждающих, что все респонденты – реальные люди, обладающие указанными характеристиками; что ни один из респондентов не заполнил анкету дважды; что параметры отбора чётко определены, и неподходящие под критерии отбора участники опроса были выявлены. Показатели, используемые для оценки качества выборки, ещё далеки от общепринятых стандартов, но они дают исследователям, работающим с опт-ин-панелями, дополнительные инструменты для такой оценки.

7.5. ДРУГИЕ ПОКАЗАТЕЛИ

Как мы отмечали вначале, два показателя, используемые для оценки репрезентативности случайных выборок в рамках модели общей ошибки исследования, – коэффициенты покрытия и коэффициенты ответа – не могут быть признаны хорошими индикаторами смещений. Исследователи, традиционно обращающиеся к случайным выборкам, должны, преодолевая это затруднение, изучать альтернативные подходы, что особенно актуально для коэффициентов ответов. Например, R-индикаторы (R-indicators) – это один из способов избежать или заместить коэффициенты ответов [Schouten, Cobben, Bethlehem, 2009]. Сёрндал [Sarndal, 2011] также обращается к этой проблеме и предлагает находить некий баланс между выборкой и критериями отбора для уменьшения ошибок неответов респондентов. В сущности, R-индикаторы и некоторые показатели, предложенные Сёрндалом и его коллегами, включают сопоставление коэффициентов ответов для подгрупп,

определяемых с помощью вспомогательных данных. Если коэффициенты ответов всех подгрупп относительно близки, для ошибок неответов остаётся мало оснований (смещения, связанные с неответами, происходят тогда, когда коэффициенты ответов коррелируют с оцениваемыми характеристиками). Индикаторы подсчитывают эту вариацию в коэффициентах ответов для разных подгрупп.

Предлагаются и более подходящие для неслучайных выборок методы. Франк и Мин [Frank, Min, 2007] предложили способ тестирования с целью оценить, может ли причинная зависимость (causal inferences) быть генерализована, если данные не являются случайной выборкой из всей совокупности. Фен [Fan, 2011] описал метод для конструирования «карт нечувствительности» (tolerance maps), обеспечивающих пользователей информацией о приемлемости разных уровней нерепрезентативности в разных исследованиях. Используя данные опросов, проводимых Louis Harris, Pew Internet и American Life Project, он продемонстрировал, как репрезентативные ответы могут быть получены из неслучайных онлайн-выборок, рекрутируемых исключительно с политически консервативных веб-сайтов. В принципе, этот новый метод вполне соответствует предлагаемым в следующем разделе идеям, которые позволяют исследователям «разрабатывать выборки, позволяющие получить ответы, которые не являются в полной мере репрезентативными, но всё же остаются в приемлемых для исследователя границах точности».

7.6. ВЫВОДЫ

Основная причина неразработанности показателей, оценивающих качество неслучайных выборок, частично кроется в отсутствии единой рамки, в которую укладывались бы все неслучайные методы, и – что, возможно, важнее – в исторически сложившемся у исследователей-практиков стереотипе восприятия этих методов как невалидных. С другой стороны, исследователи освоили различные широко применяемые показатели качества, позволяющие, по их мнению, описывать и обосновывать выводы, которые

получают из случайных выборок. Очевидно, если неслучайные методы построения выборок будут признаны валидными для исследований, потребуются аналогичные показатели и способы их измерения. Некоторую помощь в этом может оказать модель общей ошибки исследования. Ошибки ненаблюдения, особенно ошибки покрытия и неответов, создают серьёзные проблемы. Основы для неслучайных выборок редко определяются таким образом, чтобы полностью покрывать всю целевую совокупность и предоставлять каждому члену совокупности возможность быть отобранным. Показатели относительного покрытия возможны в некоторых случаях, но поскольку они в основном опираются на переменные, доступные для сопоставлений (демографические признаки), нет никаких гарантий, что они измеряют требуемые характеристики.

Наконец, гораздо важнее предложить валидные шаги по преодолению ошибок покрытия и неответов в неслучайных выборках, нежели измерить их. И сделать это нужно хотя бы для рутинных, повторяющихся практик.

Установление внешней валидности или репрезентативности посредством сравнения значений исследования с эталонными данными (такими как перепись населения, высококачественные случайные выборки или административные записи) иногда возможно, но не всегда ясно, существуют ли такие данные и не содержат ли ошибки они сами. Репликации также могут быть полезны для оценки разброса в оценках и могут проводиться относительно легко как минимум для некоторых неслучайных методов отбора.

Мы полагаем, что разумно развивать модель измерения ошибок неслучайных выборок, отталкиваясь от инструментов для случайных выборок. Когда рассматриваются ошибки наблюдения, мы вправе ожидать, что в неслучайных выборках будут ошибки, схожие с ошибками случайных выборок.

Как уже не раз говорилось в этом отчёте по разным поводам, неотъемлемый риск методов, основанных на моделировании, заключается в несоответствии предположений, на которых основана модель, изучаемым реалиям, поэтому результаты столь чувствительны к предположениям. Когда это случается, оценки смещаются, возможно, существенно. Если мы выносим

информированные суждения о качестве неслучайной выборки, все предположения, регулирующие реализацию выборки, должны быть ясно представлены, описана их эмпирическая валидация и определены используемые вспомогательные переменные. Мы призываем разрабатывать новые подходы и показатели, такие как байесовский доверительный интервал (credibility interval), но отмечаем, что перемещение этих показателей из стадии предположения в стадию принятия требует большей валидации, чем есть сейчас. Если даже исследователи разработают эмпирические показатели, подкреплённые строгой теорией, и пройдут десятилетия практики, направленной на проверку их валидности, нельзя будет выносить однозначных суждений и делить методический мир на белое и чёрное.

Транспарентность метода давно определяется AAPOR как единственный надёжный критерий качества исследования. Это особенно справедливо для исследований, основанных на неслучайных выборках. Однако до сих пор не существует общей договорённости о том, какая информация должна быть раскрыта. Коды AAPOR представляют рекомендации, разработанные для случайных выборок, но не все они подходят для таких источников, как опт-ин-панели. Стандарт Министерства государственных работ и услуг Канады [Public Works and Government Services Canada, 2008] и два стандарта ИСО (ISO 20252, ISO 26362) представляют порядок раскрытия информации для онлайн-исследований.

В итоге многообразие подходов и методов в неслучайных выборках, которые к тому же непрерывно развиваются, существенно затрудняет оценку их качества. И в большинстве случаев индивидуальные исследователи и аналитики вынуждены принимать частные решения о качестве и приемлемости тех или иных оценок, исходя из особенностей реализуемой выборки. Для того чтобы продвинуться вперёд в этом направлении, необходимо сделать устойчивыми процедуры для различных видов неслучайных выборок, например опт-ин-панелей, а также оценить полученные при помощи таких выборок результаты с использованием широкого круга методов. Если полученные результаты будут консистентны априорным теоретическим оценкам точности и стабильности показателей, поддержка таких методов возрастет в разы.

8 Неслучайные выборки

Основная причина доминирования случайных выборок на протяжении последних 60 лет – их точность, то есть они обеспечивают возможность в случае выполнения ключевых допущений получать такие оценки, которые отклоняются от истинных значений изучаемой совокупности не более чем на величину вычисляемого доверительного интервала. Всё это время дискуссии о качестве опросных данных велись в контексте статистических концептов – смещений и дисперсий (разбросов), а также попыток описания различных типов ошибок в проектируемом и реализованном дизайне опроса. Проанализировав публикации по статистике и социальным наукам за последние десятилетия об общей ошибке опроса (total survey error, TSE), Роберт Гроувз [Groves, 1989] описал базовые категории ошибок и их влияние на смещение и разброс и связал это со стоимостью проводимого исследования.

Как показано в предыдущем разделе, модель общей ошибки опроса представляет отличную линзу для рассмотрения источников ошибок в вероятностных исследованиях с потенциальной возможностью её применения и для неслучайных выборок. Однако Гроувз и Лайберг [Groves, Lyberg, 2010] сходятся в том, что модель общей ошибки опроса должна быть помещена в более широкую рамку качества исследования. Во-первых, важнейшую роль играет стоимость, или цена проекта, которая должна рассматриваться в предельно прагматичном ключе. Во-вторых, оценка результатов исследования должна учитывать, как полученные данные будут применяться, то есть отвечать таким критериям, как соответствие целям (fit for purpose) или пригодности для использования (fitness for use).

8.1. ИЗМЕНЕНИЕ ОПРЕДЕЛЕНИЯ КАЧЕСТВА

В своей книге Бимер и Лайберг [Biemer, Lyberg, 2003] связывают изменение определений качества опросных данных с более широкой революцией качества 1980–1990-х годов. В производственных спецификациях качество определялось как нечто «свободное от дефектов» или «соответствующее спецификациям». В результате движения «комплексного управления качеством» (Total Quality Management, TQM) произошел отказ от прежних представлений, согласно которым качество непосредственно связано с субъективными потребностями потребителя и с тем, как продукт будет использоваться. «Качество должно отвечать запросам потребителей» [Deming, 1982]. Введение в научный оборот концепта «соответствие целям» часто связывают с именем Юрана [Juran, 1992]. Под ним понимались особенности использования продукта и цена, которую покупатель готов за это заплатить. Именно эти факторы играют важную роль в процессе проектирования и выступают базовой частью концепта качества.

ИСО, Международная организация по стандартизации[28], считает своей миссией разработку и распространение международных стандартов на продукты и услуги. Работая через сеть национальных институтов из 162 стран, ИСО разработала более 19 000 глобальных стандартов «для утверждения желательных характеристик продуктов и услуг, таких как качество, экологичность, безопасность, надёжность, эффективность и заменяемость, и всё это при условии резонных экономических затрат».

Так, сейчас во всём мире понятие качества не является абсолютным, а, напротив, определяется в контексте ожиданий покупателей – целей, для которых были приобретены продукты или услуги, и соответствия этим целям.

---

[28] «ИСО (ISO), Международная организация по стандартизации, основана в 1947 году и с тех пор опубликовала более 19 500 международных стандартов, которые распространяются почти на все аспекты технологии и бизнеса, от безопасности пищевых продуктов до компьютеров, а также сельского хозяйства и здравоохранения» (см. подробнее: http://www.iso.org/iso/ru/home/about.htm) — *Прим. перев.*

8.2. КОНЦЕПТ КАЧЕСТВА В МАССОВЫХ ОПРОСАХ

Имя Деминга в обыденном сознании чаще всего связывается с так называемой революцией качества и движением комплексного управления качеством – сначала в послевоенной Японии и затем в США. Среди исследователей он чаще воспринимается как знаменитый статистик, который в начале 1940-х внедрял выборочные процедуры в Бюро переписи США [Aguayo, 1990]. Именно в эти годы Деминг пришел к выводу, что точность не должна быть единственным критерием для оценки данных опросов [Deming, 1944]. Не менее значимой является «полезность», под которой он понимал «помощь в формировании рациональных оснований для действий» [Deming, 1944, p. 369].

В литературе по выборке и массовым опросам периодически поднимаются вопросы о соотношении между качеством и полезностью данных. Например, Садмен [Sudman, 1976], сопоставляя качество выборки с тем, как информация будет использоваться, выделял два полюса. Первый он называл «зондирующий (разведывательный) сбор данных», основная цель которого – формулировка первоначальных гипотез для дальнейших исследований. Но втором полюсе он располагал большие национальные исследовательские проекты, направленные на сопровождение государственных решений и программ. По Садмену, второй полюс требует большей точности измерения, чем первый. Соответственно, место исследования на этой шкале необходимо определять при помощи теста на пригодность для использования, одной из важных составляющих которого выступает точность оценки. Садмен предложил 10 особых примеров исследований, в которых дизайны выборки представляли разные степени компромиссов, в основном определяемых стоимостью или сложностью реализации. Однако практически во всех случаях акцент делался на том, будет ли потеря точности из-за этого компромисса существенно влиять на полезность оценок.

В том же духе высказывался Киш: «Статистический дизайн всегда включает компромиссы между желаемым и возможным» [Kish, 1987, p. 1]. Для Киша такие компромиссы определяются текущими возможностями и ресурсами, привлекаемыми для достижения целей исследования. Он выделял три основные категории, или области, в которых обычно идут на компромиссы:

1 — репрезентативность, связанная главным образом с неполной основой выборки и низким уровнем неответов; 2 — рандомизация, означающая поиск путей для учёта эффекта возмущающих (confounding) переменных; 3 — реализм представления данных, или степень соответствия переменных конструктам, которые они описывают. Киш предложил упорядоченный список 10 исследовательских дизайнов, на одном конце которого располагались исследования, где важна репрезентативность, а на другом – эксперименты, в которых критична рандомизация. Исследования, основанные на наблюдениях (observational studies), где на первый план выходит критерий реализма, он расположил посредине. Этот список исследовательских дизайнов позволяет различить то, что Киш называет «счётными», или описательными исследованиями, с одной стороны, и «экспериментальными» или аналитическими, — с другой.

Гроувз придерживается схожей точки зрения [Groves, 1989], когда обращается к различиям в потребностях тех, кого он называет «описателями» (describers), и тех, кого он называет «разработчиками моделей» (modelers). Первым необходимы данные опросов, полностью отражающие совокупность, которую они хотят описать, и для них особое значение приобретают ошибки ненаблюдения[29] . Вторые заинтересованы в данных, которые наиболее полно представляют концепты, позволяющие тестировать их теории, и гораздо меньше озабочены ошибками покрытия и неответов. Правительственные агентства чаще относятся к описателям, поскольку их основной интерес заключается в точной оценке некоторых специфических характеристик целевых групп населения. Научные сотрудники и иногда маркетологи чаще занимают позицию разработчиков моделей. Они интересуются тем, как личные характеристики влияют на голосование за того или иного кандидата или выбор того или иного продукта. Описатели и разработчики моделей имеют свои собственные области интересов, а поэтому и разные представления о том, что определяет высокий уровень качества данных.

[29] Ошибки ненаблюдения (non-observational errors) – общее название для ошибок покрытия и неответов. – *Прим. перев.*

Омюрхатай также связывает потребности тех, кто будет анализировать данные, и причины сбора данных с качеством: «Концепт качества, и более того – концепт ошибки, может быть удовлетворительно определён лишь в том контексте, в котором проводилось исследование. Смысл ошибки может меняться в той же степени, в какой меняются контекст и цели исследования. Задача переформулируется в терминах целей и теоретической рамки исследователя, а не установленных условных (псевдообъективных) критериев. Это сразу снимает потребность в определении истинных значений концептов в некотором абсолютном смысле и требует концентрации внимания на потребностях, согласно которым собирались данные» [O'Muircheartaigh, 1997, p. 1].

8.3. СООТВЕТСТВИЕ ЦЕЛЯМ В ГОСУДАРСТВЕННЫХ СТАТИСТИЧЕСКИХ АГЕНТСТВАХ

Государственные статистические агентства представляют собой классический пример описателей, для которых точность – основной атрибут качества, зачастую вследствие того, что полученные оценки играют ключевую роль в принятии финансовых и политических решений. Более того, значительная часть требований статистических агентств включает дополнительные критерии качества, с учётом которых могут приниматься решения о подходящем дизайне исследования. В 2002 году Статистическое агентство Канады опубликовало руководства, которые определили шесть элементов качества для реализуемых агентством статистических программ [Statistics Canada, 2002, p. 4.]:

- *релевантность* – степень соответствия целям агентства или политическим целям;
- *точность* – степень, с которой оценки корректно измеряют то, что они должны измерить, учитывая допустимый диапазон ошибки;
- *своевременность* – вероятность того, что данные будут доступны тогда, когда это требуется для сопровождения принимаемых решений;

- *доступность* – наличие данных в форме, пригодной для тех, кому они нужны;
- *интерпретируемость* – лёгкость понимания пользователями дизайна исследования и процедур сбора данных для вынесения суждений об их применимости;
- *согласованность* – степень соответствия данных статистической программе и совместимости с другими аналогичными по времени и географии данными.

Австралийское бюро статистики разработало схожую концепцию [Australian Bureau of Statistics, 2009]. В ней добавлен седьмой элемент качества, который называется «институциональная среда». Он обусловлен брендом агентства или организации, с которыми ассоциируется статистический продукт, – в какой степени это агентство или организация воспринимается как объективное, независимое и защищающее конфиденциальность участников исследования.

Другие агентства, такие как Евростат [Eurostat, 2003], Статистическое агентство Швеции [Rosen, Evers, 1999], Международный валютный фонд [Carson, 2001] и Организация экономического сотрудничества и развития [OECD, 2002], придерживаются концепций, которые представляют собой вариации перечисленных принципов.

8.4. СООТВЕТСТВИЕ ЦЕЛЯМ В МАРКЕТИНГОВЫХ ИССЛЕДОВАНИЯХ

Маркетинговые исследователи ведут себя то как описатели, то как разработчики моделей. Мониторинговые исследования, в которых регулярно измеряются такие показатели, как удовлетворённость продуктом или стабильность его использования, в некоторых аспектах весьма схожи с исследованиями государственных статистических агентств. Медиаизмерения, затрагивающие телесмотрение или чтение периодики, – другой пример, где весьма желательны точные оценки. В этих примерах, как и в описанных

выше исследованиях государственных агентств, применяется критерий пригодности для использования (fitness for use), разве что больший акцент иногда делается на стоимости, своевременности и доступности. В других случаях маркетинговые исследователи более схожи с разработчиками моделей, которым собранные данные важны в первую очередь для тестирования статистических моделей, представляющих, например, как связаны личностные характеристики и свойства продукта. Это нужно для успешного продвижения продукта там, где другие потерпели неудачу.

Соответственно, маркетинговые исследователи явно или неявно принимают концепт пригодности для использования (fitness for use). Особенно это очевидно в специфическом контексте опт-ин-панелей [Bain, 2011]. Однако слишком мало написано об особенностях общей рамки, которой придерживаются государственные статистические агентства. Европейское общество по изучению общественного мнения и рынка (ESOMAR) регулярно публикует руководства по выбору специфических методов, соответствующих целям исследования. Например, в их «26 вопросах для правильной покупки онлайн-выборок» [ESOMAR, 2008] предполагается, что при выборе провайдера онлайн-выборки следует обращать внимание на семь моментов: профайл компании, источники формирования выборки, методы рекрутинга, практики управления панелью, соответствие нормам индустрии и применяемому в ней законодательству, сотрудничество с другими поставщиками выборок и методы оценки качества и валидности данных. В руководстве «36 вопросов для помощи в проведении неврологических исследований» [ESOMAR, 2011] применяется несколько отличный набор критериев. В соответствии с заглавиями оба документа представляют вопросы, которые должны быть заданы, описывают, почему каждый вопрос важен, но оставляют на усмотрение исследователя оценку влияния каждого ответа на качество выбранного дизайна исследования.

Смит и Флетчер предложили более понятную рамку. Они начали с предпосылки, что «зачастую присутствует неизбежный зазор между качеством первичных данных и требуемых от маркетинговых исследований заключений» [Smith, Fletcher, 2004, p. 61]. Основная задача аналитика – преодолеть разрыв между идеальным и реальным, между данными, которые мы хотим

получить и которые мы реально получаем. Для решения этой задачи Смит и Флетчер адресуют аналитикам восемь методологических вопросов, большинство из которых можно проинтерпретировать в терминологии ошибок, описанных в стандартной модели общей ошибки исследования (total survey error model). Один из восьми вопросов заключается в оценке соответствия дизайна исследования поставленной цели, что определяется авторами как компромисс между пятью ключевыми переменными:

- требуемая точность измерения;
- глубина или уровень детализации в данных;
- практические и этические ограничения;
- дата предоставления данных (время, отпущенное на исследование);
- бюджет.

8.5. ВЫВОДЫ

Постоянные компромиссы в исследовательском дизайне – общее место для всех без исключения исследований во всех возможных секторах исследовательской индустрии. Приверженцы случайных выборок должны определиться, как им следует относиться к недопустимо высокому уровню неответов. Огромный рост популярности опт-ин-панелей должен сопровождаться осознанием чрезвычайно больших ошибок покрытия и отбора и готовностью мириться с ними. Подобные компромиссы получают всё большее распространение на практике, но они редко имеют осознанную ориентацию на принцип пригодности для использования. Даже при самых идеальных условиях, в которых привычные ограничения бюджета, времени и реализуемости намеченного плана (feasibility) не требуют компромиссных решений, степень соответствия полученных результатов запросам заказчиков исследований сейчас воспринимается в качестве ключевого параметра любого исследовательского дизайна. Можно лишь повторить слова Гроувза и Лайберга: «<…> статистика не имеет никакого значения без определения способов её применения» [Groves, Lyberg, 2010, p. 873].

Применение идеи «соответствия цели» для определения качества и как ключевого принципа разработки дизайна исследования имеет некоторую историю. В кратком её представлении Бимер и Лайберг приходят к пониманию соответствия цели как основного определения качества исследования. Они задают три ключевых параметра качества: точность, своевременность и доступность. Для них исследование может быть признано качественным только тогда, когда оно выполнено «настолько точно, насколько это необходимо для достижения поставленных целей, данные подготовлены к указанному сроку и доступны тем, для кого исследование проводилось» [Biemer, Lyberg, 2003, p. 13]. Множество государственных статистических агентств усложняют и развивают это базовое представление о качестве, однако в нашем представлении подобное усложнение принципов зачастую труднореализуемо на практике.

Большинство принципов, определяющих качество исследования, построено на различиях между потребностями описателей и разработчиков моделей. Потребности первых довольно хорошо укладываются в представление об общей ошибке опроса, разделяемое многими исследователями. Описатели более всего озабочены снижением ошибок покрытия и неответов как способом повышения репрезентативности и точности. Они чаще всего выбирают вероятностные дизайны выборок.

Разработчики моделей, напротив, больше внимания уделяют измерению всех концептов, которые, по их мнению, играют важную роль в объяснении поведения, являющегося предметом их исследований. Их интерес заключается скорее в выявлении взаимосвязи широкого набора характеристик, нежели в точном измерении значений этих характеристик в изучаемой совокупности. Они не игнорируют ошибки ненаблюдения (non-observation), но, как правило, предполагают отсутствие значимой связи между изучаемыми явлениями, зависимыми переменными и вероятностью отбора конкретного человека или его готовности отвечать. Другими словами, их основная задача заключается в достижении внутренней валидности. Разработчики моделей наиболее часто обращаются к неслучайным выборкам как к менее затратным, чем те, которые предпочитают описатели.

Полезно осознавать, что описанные различения не представляют дихотомии, а лишь задают полярные значения одного континуума. Упорядоченный список исследовательских дизайнов Киша, возможно, лучшая рамка для того, чтобы согласовывать цель и необходимую точность исследования с практическими ограничениями по срокам, бюджету, доступности и общей реализуемости исследовательского проекта.

В некоторых случаях выбор может сводиться к неслучайной выборке или вовсе к отказу от опроса. Если сохраняется некоторый уровень доверия к тому, что допущения используемой модели достаточны для целей исследования, реализация неслучайной выборки вполне оправданна. Если этого нет, оценки, полученные по неслучайной выборке, могут приводить к настолько плохим решениям, что лучше вовсе отказаться от проведения опроса.

9 Заключение

Надеемся, нам удалось осветить основные проблемы, с которыми сталкиваются работающие с неслучайными выборками исследователи. Вследствие непроработанности тематики неслучайных выборок наши усилия могут расцениваться как не слишком успешные, а описания – как недостаточно полные, если сравнивать с более разработанной сферой массовых опросов. Многие представленные выше методы давно применяются в разных дисциплинах, но до сих пор остаются неизвестными широкому кругу исследователей опросной индустрии. Мы полагаем, что наш отчёт послужит началом широкого освещения этих методов, для чего и обобщаем ниже наиболее важные положения.

В отличие от ситуации со случайными выборками, не существует общей теоретической рамки, которая адекватно представляла бы все способы неслучайных отборов. Хотя есть множество разновидностей случайной выборки, все они опираются на общий теоретический базис – для оценки характеристик совокупности надо построить относительно большую выборку, в которой известен шанс отбора каждой единицы, и взвесить единицы выборки обратно вероятностям их отбора (с корректировкой для пропущенных данных, если это необходимо). В сравнении с этим мы располагаем большим разнообразием методов построения неслучайных выборок, различающихся как по способам построения, так и по формированию оценок. Статистические свойства и практическая реализация этих методов весьма разнообразны. Соответственно, неслучайные выборки – это коллекция методов, и чрезвычайно сложно, если вообще возможно, описать общие свойства всех подходов к построению неслучайных выборок. Прежде чем обращаться к неслучайным методам, исследователь должен

понимать их специфические свойства и ограничения, особенно это относится к ограничениям на статистические выводы, которые могут относиться ко всей совокупности.

Может оказаться плодотворным представление о различных подходах к построению неслучайных выборок как некотором континууме ожидаемой точности оценок. На одном конце этого континуума располагаются неконтролируемые конформные выборки, оценки которых основаны на допущении, что проведён случайный отбор респондентов из совокупности. Такие исследования обычно имеют мало оснований для подобных допущений или совсем их не имеют. Выводы, получаемые из таких выборок, довольно рискованны. На другом конце континуума расположены методы, в которых респонденты отбираются в соответствии с критерием, согласованным с предметом исследования, и результаты корректируются с использованием переменных, которые связаны (коррелируют) с ключевыми результатами. Когда имеются данные, подтверждающие используемые предположения, выводы из таких выборок подвержены меньшему риску[30]. Основная сложность возникает при размещении исследования между этим двумя крайностями.

Для социальных и маркетинговых исследований это огромная неизученная территория. Оценка риска зачастую зависит как от знания об изучаемой совокупности, так и от технических особенностей применяемых методов. Понимание, насколько разные значения изучаемых характеристик могут наблюдаться в совокупности и как предполагается использовать полученные оценки, чрезвычайно критично для оценки рисков, сопровождающих статистические выводы. Наиболее распространённая ошибка в изучении человеческого поведения и установок – это допущение о высокой гомогенности (схожести людей между собой). Такая ошибка приводит к смещениям и недооценке разброса в получаемых результатах. Качество контроля выборки и корректировки данных, возможно, оценить ещё сложнее. Мы полагаем, что определённый контроль за отбором очень важен; случайная выборка

[30] На практике можно говорить лишь об единичных случаях (включая случайные выборки), которые исходя из такого критерия могут быть отнесены к исследования с низким риском.

представляет пример наиболее контролируемого механизма отбора. Рубин [Rubin, 2008] приводит аналогичные суждения в несколько ином контексте. Важность контроля выборки – основание для другого заключения (ниже), суть которого в том, что выравнивание (matching) выборки является наиболее многообещающим подходом для неслучайных опросов. Совмещение контроля выборки с использованием хороших вспомогательных переменных на стадии корректировки данных должно обеспечить меньшую рискованность статистических выводов из неслучайных выборок.

Транспарентность (прозрачность) – неотъемлемая методическая сущность. Для неслучайных выборок гораздо сложнее описывать применяемые методы отбора респондентов, сбора данных и производства статистических выводов, чем для случайных выборок. Вновь напомним, что не существует единого метода получения неслучайной выборки, и различия в методах могут быть значительными. Поэтому точное описание методов и предположений критично для понимания применимости получаемых оценок. Когда предположения чётко определены и имеются данные, подтверждающие их справедливость, пользователь может произвести информированную оценку (informed assessment) рисков, связанных с полученными из исследования выводами. Такая оценка не является простым действием, и чтобы неслучайные выборки в целом были включены в научный оборот, требуется ясное их описание. Методологии чёрного ящика должны быть вскрыты и стать полностью транспарентными. Частная информация не всегда может быть раскрыта, но методология должна быть прозрачной, а ключевые предположения – описанными и систематизированными. В частности, большинство онлайн-исследований не отвечают этому требованию – не предоставляют информацию, адекватно описывающую их методологию. Это следует изменить.

Производство статистических выводов и для случайных, и для неслучайных исследований требует определённого доверия к тем допущениям, на которых основаны модели расчётов. Эти допущения должны быть чётко описаны, с максимальной полнотой должны быть описаны и эффекты, которые возникают в результате этих допущений и могут влиять на точность оценок. Хотя наш отчёт посвящён неслучайным выборкам,

разработчикам случайных выборок также следует осознавать, что невозможность получения 100%-ного уровня покрытия и коэффициента ответов означает, что и случайные выборки часто опираются на ряд допущений.

Наиболее перспективные неслучайные методы основаны на моделях, в которых проблемы, связанные со статистическим выводом, преодолеваются как на этапе отбора, так и на этапе оценки. Хотя случайный отбор остается базовым стандартом для производства статистического вывода, это не только не единственный, но даже не самый распространённый способ в общей статистической практике. Другие подходы, которые в сфере массовых опросов принято называть подходами, основанными на моделировании [Valliant, Royall, Dorfmann, 2000], используются во многих неопросных технологиях. В этих подходах обычно предполагается, что ответы возникают в соответствии со статистической моделью (например, все наблюдения имеют одни и те же среднее и дисперсию) и что применимость и адекватность этой модели можно улучшить за счёт использования важных вспомогательных переменных. Как только модель сформулирована, для получения статистических выводов об оцениваемых параметрах совокупности применяются стандартные статистические процедуры оценивания, такие как вероятностные, или байесовские техники. Безусловно, допущения упрощённых моделей, игнорирующие сложность изучаемых явлений, не могут быть признаны адекватными. Даже с методами, основанными на моделировании (model-based methods), получение статистического вывода остается сложной задачей.

Одна из причин нечастого использования в массовых опросах методов, основанных на моделировании, заключается в том, что разработка релевантных моделей и тестирование заложенных в них допущений весьма сложны и затратны по времени, к тому же требуют хорошей статистической квалификации. Допущения должны быть проверены для всех ключевых оценок – модель, хорошо работающая с одними оценками, может давать сбои с другими. Один из наиболее важных атрибутов случайной выборки состоит в том, что существует стандартный подход, позволяющий получить широкий спектр оценок, многие из которых часто востребованы. Хотя для случайной выборки требуются допущения, позволяющие учиты-

вать пропущенные данные, эти допущения относительно стандартны, и подходы к корректировке выборки становятся рутинной практикой. Достижение простоты, присущей случайным выборочным методам в производстве множественных оценок, весьма проблематично для методов неслучайного отбора. В сущности, даже при более дешёвом сборе данных в неслучайных выборках по сравнению со случайными менее контролируемый метод отбора требует больших усилий на стадии анализа.

Соответствие цели – важнейший концепт для оценки качества опросных данных, но его применение в дизайне исследования требует дальнейших разработок. При описании комплексного моделирования в отчёте Национальной академии наук указывалось, что «уровень строгости выполняемой работы должен быть соразмерен важности и потребностям её практического приложения и контексту принятия решения. Некоторые приложения сопровождаются принятием важных решений, и поэтому требуют значительных <…> усилий; другие – нет» [National Academy of Science, 2012, p. 96]. Существует консенсус между всё возрастающим числом национальных статистических агентств по ключевым элементам, которые позволяют формировать качественную основу для поддержки, развития и реализации дизайна исследований, соответствующего поставленным целям. В основном предполагается установление баланса между такими ключевыми элементами, как релевантность, точность, своевременность, доступность, интерпретируемость и согласованность. Следующий логический шаг – переход от теоретической рамки к согласованным с ней практическим действиям. Неслучайные выборки должны оцениваться по аналогичным принципам.

Методы построения выборки, используемые в опт-ин-панелях, с течением времени значительно эволюционировали, поэтому исследователь, которому необходимо оценить валидность получаемых данных, должен скорее обращать внимание на методы отбора, чем на саму панель. Наблюдается тенденция сваливать в одну кучу все онлайн-выборки, сформированные из опт-ин-панелей, как будто опт-ин-панели являются методом отбора. Но это не так. Пользователи опт-ин-панелей могут применять различные методы отбора, процедуры сбора данных и техники оценивания. Исследовательские оценки прежних методов построения неслучайных

выборок из панелей малоприменимы для анализа текущих подходов. Следует перенести фокус исследований с опт-ин-панелей на изучение разных стратегий построения выборки и получения оценок, которые можно реализовать на опт-ин-панелях.

Если неслучайные выборки станут шире использоваться в практике массовых опросов, потребуется более согласованная теоретическая рамка и сопутствующий набор измерителей для оценки их качества. Одно из ключевых достоинств случайного отбора – это комплект готовых измерителей и конструктов (таких как общая ошибка исследования), которые определяют понимание качества и позволяют выявлять источники ошибок. Применение этого набора для оценки неслучайных выборок неэффективно из-за иных оснований отбора. Общепризнана острая необходимость в исследованиях, нацеленных на развитие показателей качества, в том числе оценки смещений и точности полученных на неслучайных выборках значений.

Неслучайные выборки хорошо себя зарекомендовали в электоральных опросах, но нет столь же определённых свидетельств их качества в других сферах, в том числе в комплексных обследованиях с целью изучения различных феноменов. Исследования, нацеленные на получение ограниченного числа оценок по заданному набору результатов, требуют контроля за малым набором сопутствующих переменных. Повторяющиеся замеры, как в мониторинговых исследованиях, и доступность внешних эталонных данных (таких как результаты выборов) позволяют проводить эксперименты и могут упрощать разработку моделей. Однако во многих исследованиях такие возможности отсутствуют. Часто оценивается множество переменных из широкого круга предметных областей, что требует привлечения большого набора сопутствующих переменных (ковариат).

Неслучайные выборки могут быть пригодны для производства статистических выводов, но валидность этих выводов зависит от уместности допущений, заложенных в модель, и от того, насколько отклонения от допущений влияют на конкретные оценки. На протяжении всего отчёта мы проводим мысль, что для любого метода построения неслучайных

выборок необходимо развивать теоретическую основу, за которой должна следовать эмпирическая оценка состоятельности этого метода. В этой оценке должны проверяться пригодность допущений в различных случаях и для разных статистических расчётов. В отчёте подчеркивается, что выравнивание выборки – это один из методов, уже имеющих сконструированную для оценочных исследований теоретическую основу, которая должна быть модифицирована и адаптирована под требования массовых опросов. Некоторые исследователи приступили к решению этой задачи. Методы постопросных корректировок, применяемые в неслучайных выборках, в большинстве случаев копируют подходы, применяемые для случайных выборок. Хотя иногда они могут быть пригодны и эффективны, требуется дополнительное рассмотрение механизмов смещения при отборе. Мы считаем, что повестка по развитию метода должна включать эти требования.

АНГЛО-РУССКИЙ СЛОВАРЬ ТЕРМИНОВ

absorption rate	коэффициент абсорбции, коэффициент поглощения (для опт-ин-панелей)
access panel	аксесс-панель (панель людей, имеющих доступ в интернет)
accuracy	точность
adjustment factor	корректирующий (поправочный) коэффициент
allocation	размещение, распределение (объектов в выборке)
attrition rate	коэффициент износа (панели)
auxiliary data / variables	вспомогательные (сопутствующие, дополнительные) данные / переменные
Bayesian method	байесовский метод
benchmark	эталон
bias	смещение
break off rate	коэффициент прерываний
calibration adjustment	калибровочная корректировка
case-control study	исследования методом «случай-контроль»
clinical trial	клиническое испытание, клинический эксперимент
cluster sample	кластерная (гнездовая) выборка
cobort study	когортное исследование
completion rate	коэффициент заполнения (анкеты)
confidence interval	доверительный интервал
controlled variables	контролируемые переменные
convenience sampling	конформная выборка
covariate	ковариата (сопутствующая, вспомогателная, дополнительная переменная)
coverage	покрытие, охват
coverage error / bias	ошибка / смещение покрытия
coverage ratio	коэффициент покрытия
credibility interval, credible region	байесовский доверительный интервал

design-based approach	подход, основанный на проектировании [выборки]
disconnected graph	несвязный граф
disturbing variables	возмущающие переменные
effective sample size	эффективный размер выборки
eligibility criteria	критерий пригодности, приемлемости, соответствия
eligibility rate	коэффициент пригодности, приемлемости, соответствия
error of non-observation	ошибки ненаблюдения (отсутствия наблюдения)
error of observation	ошибки наблюдения
estimation error	ошибка оценки, ошибка оценивания
evaluation research	оценочное исследование
exclusion bias	ошибка невключения (для неслучайных выборок, в отличие от ошибки покрытия для случайных выборок)
explanatory variable	объясняющая переменная
external validity	внешняя валидность
fitness for use	пригодность для использования
frontier sampling	пограничная выборка
generalized regression weighting (GREG)	обобщённое регрессионное взвешивание
homophily	гемофильность
incidence study	инцидентное (изучающее историю заболеваний) исследование
intercept survey	перехватывающее исследование
internal validity	внутренняя валидность
jackknife	метод «складного ножа», метод «выкидного ножа»
judgmental sample	экспертная выборка
link-tracing network sample	сетевая выборка с отслеживанием связей
longitudinal study	лонгитюдное (панельное) исследование
mall-intercept	перехватывающие опросы в торговых центрах

mean square error	среднеквадратическая ошибка
missing at random	несистематические случайные пропуски
model-based approach	подход, основанный на моделировании
multiple frame sample	многоосновная выборка
multiplicity sampling	многократный отбор
network sampling	сетевая выборка, сетевой отбор
non-observational errors	ошибки ненаблюдения
non-probability	неслучайная, невероятностная выборка
sample nonresponse	неответы (отсутствие ответов)
nonresponse error / bias	ошибка / смещение неответов
observational study	исследование, основанное на наблюдениях
one-to-one matching	выравнивания один к одному
opt-in panel	опт-ин-панель (панель согласившихся принимать участие в онлайн-опросах)
participation rate	коэффициент участия
post hoc adjustments	post hoc корректировки
poststratification	постстратификация, стратификация после отбора
predictor variable	предсказывающая переменная (предиктор)
probability sample	случайная, вероятностная выборка
profile survey	опросный профиль
propensity score adjustment (PSA)	корректировка по степени склонности
pseudo-replication	псевдорепликация, псевдо-тиражирование, псевдоповторение выборки
purposive selection / sample	целевой отбор / выборка
random digit dial, RDD	случайный набор номера
randomized controlled laboratory study	рандомизированное контролируемое лабораторное исследование
randomized controlled trial	рандомизированное контролируемое испытание
randomized cross-over trial	рандомизированное перекрестное испытание
randomized trial	рандомизированный экспериментальный план, случайный эксперимент

randomized variable	случайная переменная
replication	репликация, тиражирование, повторение (выборки)
resampling	повторная выборка
respondent driven sample, RDS	выборка, управляемая респондентами
respondent's degree	ранг респондента (число его контактов в социальной сети)
response rate	коэффициент ответов
river sample	потоковая выборка
robo-calls	роботизированные телефонные опросы
sample	выборка
sample design	дизайн выборки, конструкция выборки
sample frame	основа выборки, выборочный фрейм
sample matching	выравнивание (сопряжение) выборки (приведение в соответствие с параметрами совокупности)
sample unit (primary — PSU, secondary — SSU)	единица отбора (первичная, вторичная)
satisficer	минимизирующий усилия (см. satisficing)
satisficing	минимизация усилий (на заполнение анкеты): этот термин является комбинацией двух терминов: satisfy (удовлетворять) и suffice (быть достаточным)
screening completion rate	коэффициент завершения скрининга
seeds	первоисточники, стартовые точки (для выборки методом снежного кома)
selection bias / error	смещение / ошибка отбора
significance test	оценка уровня значимости
simple random sample (SRS)	простая случайная выборка
snowball sample	выборка методом снежного кома
standard error	стандартная ошибка
standard deviation	стандартное отклонение
statistical inference	статистический вывод

stratification	стратификация (расслоение, типическое районирование)
stratified sample	стратифицированная (расслоенная, районированная) выборка
stratum (мн:strata)	страта (слой, типический район)
systematic sample	систематическая выборка
target population	изучаемая совокупность, целевая совокупность, генеральная совокупность
targeted sample	целевая выборка
time-location sample	выборка по времени и месту
trial	испытание, эксперимент, проба
total survey error, TSE	общая ошибка исследования
variance	дисперсия, разброс
view rate	коэффициент рассмотрения
volunteer sample	выборка добровольцев
weight	вес
weight adjustment	весовая корректировка, весовая поправка
weighting	взвешивание

ЛИТЕРАТУРА

1. AAPOR (American Association for Public Opinion Research). 2012. “Understanding a ‘Credibility Interval’ and How it Differs from the ‘Margin of Sampling Error’ in a Public Opinion Poll”. Downloaded from http://aapor.org/AM/Template.cfm?Section=Understanding_a_credibility_interval_and_how_it_differs_from_the_margin_of_sampling_error_in_a_publi&Template=/CM/ContentDisplay.cf m&ContentID=5475
2. AAPOR. (American Association for Public Opinion Research). 2009. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. Revised 2011.
3. Abate, Tom. 1998. “Accuracy of Online Surveys May Make Phone Polls Obsolete”. The San Francisco Chronicle, D1.
4. Alvarez, R. Michael, and Carla VanBeselaere. 2005. “Web-Based Surveys”. The Encyclopedia of Measurement. California Institute of Technology. http://www.mta.ca/~cvanbese/encyclopedia_new2.pdf.
5. Asur, Sitaram, and Bernardo A. Huberman. 2010. “Predicting the Future with Social Media”. Downloaded from http://arxiv.org/pdf/1003.5699v1.
6. Australian Bureau of Statistics 2009. ABS Data Quality Framework, Downloaded from http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0 on 4/30/2013
7. Baker, Reg, Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Groves, Courtney Kennedy, Jon Krosnick, Paul J. Lavrakas, Sunghee Lee, Michael Link, Linda Piekarski, Kumer Rao, Randall K. Thomas, and Dan Zahs. 2010. “AAPOR Report on Online Panels”. Public Opinion Quarterly. 74(4): 711–81.
8. Banks, David. 2011. “Reproducible Research: A Range of Response Statistics”. Politics, and Policy, 2, DOI: 10.2202/2151–7509.1023.
9. Berinsky, Adam J. 2006. “American Public Opinion in the 1930s and 1940s: The Analysis Of Quota- Controlled Sample Survey Data”. Public Opinion Quarterly 70(4):499–529.
10. Bernhardt, Annette, Ruth Milkman, Nik Theodore, Douglas Heckathorn, Mirabai Auer, James DeFilippis, Ana Luz González, Victor Narro, Jason Perelshteyn, Diane Polson, and Michael Spiller. 2009. “Broken Laws,

Unprotected Workers: Violations of Employment and Labor Laws in America's Cities". Downloaded from http://www.nelp.org/page/brokenlaws/BrokenLawsReport2009.pdf?nocdn=1.

11. Berzofsky, Marcus E., Rick L. Williams, and Paul P. Biemer. 2009. "Combining Probability and Non- Probability Sampling Methods: Model-Aided Sampling and the O*NET Data Collection Program". Survey Practice August: 1-5. http://surveypractice.files.wordpress.com/2009/08/berzofsky.pdf.
12. Bethlehem, Jelke. 2010. "Selection Bias in Web Surveys". International Statistical Review 78(2):161–88.
13. Bethlehem, Jelke, and Silvia Biffignandi. 2012. Handbook of Web Surveys. Hoboken, New Jersey: John Wiley & Sons Inc.
14. Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. Handbook of Nonresponse in Household Surveys. Hoboken, NJ: John Wiley & Sons.
15. Biemer, Paul P., and Lars E. Lyberg. 2003. Introduction to Survey Quality. New York: Wiley.
16. Biernacki, Patrick, and Dan Waldorf. 1981. "Snowball Sampling: Problem And Techniques Of Chain Referral Sampling". Sociological Methods and Research 10(2):141–63.
17. Birnbaum, Zygmunt William, and Monroe G. Sirken. 1965. "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates". National Center for Health Statistics. Vital and Health Statistics 2(11).
18. Blair, Johnny, and Frederick Conrad. 2011. "Sample Size for Cognitive Interview Pretesting". Public Opinion Quarterly 75(4):636–58.
19. Brick, J. Michael. 1990. "Multiplicity Sampling in an RDD Telephone Survey". Sampling Design Issues section of the American Statistical Association 296–301.
20. Brick, J. Michael. 2011. "The Future Of Survey Sampling". Public Opinion Quarterly 75(5):872–88.
21. Brick, J. Michael, and Douglas Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional". The ANNALS of the American Academy of Political and Social Science 645(1):36–59.
22. Bryson, Maurice C. 1976. "The Literary Digest Poll: Making of a Statistical Myth". American Statistical Association 30(4):184–85.
23. Callegaro, Mario, and Charles DeSogra. 2008. "Computing Response Metrics for Online Panels". Public Opinion Quarterly 72(5):1008–32.

24. Campbell, Donald T., and Julian C. Stanley. 1963. Experimental And Quasi-Experimental Designs For Research. Chicago, IL: Rand-McNally.
25. Carlson, Robert G., Jichuan Wang, Harvey A. Siegal, Russel S. Falck, Jie Guo. 1994. “An Ethnographic Approach To Targeted Sampling: Problems And Solutions In AIDS Prevention Research Among Injection Drug And Crack-Cocaine Users”. Human Organization 53:279–86.
26. Centers for Disease Control and Prevention. 2005. “Statistical Methodology of the National Immunization Survey 1994-2002”. Vital and Health Statistics Series 2, 138.
27. Chang, Linchiat, and Jon A. Krosnick. 2009. “National Surveys via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality”. Public Opinion Quarterly 73(4):641–78.
28. Chui. Michael, Markus Löffler, and Roger Roberts. 2010. “The Internet of Things”. McKinsey Quarterly.
29. Clinton, Joshua D. 2001. “Panel Bias from Attrition and Conditioning: A Case Study of the Knowledge Networks Panel”. Paper presented at 56th Annual Conference of the American Association for Public Opinion Research, May, Montreal, Canada.
30. Cochran, William G. 1965. “The Planning of Observational Studies of Human Populations”. Journal of The Royal Statistical Society Series A, 128(2):234–66.
31. Coleman, James S., 1958. “Relational Analysis: The Study of Social Organizations with Survey Methods”. Human Organization 17:28–36.
32. Copas, John B. and H. G. Li. 1997. “Inference for Non-random Samples”. Journal of the Royal Statistical Society Series B, 59:5-95.
33. Cornfield, Jerome. 1971. “The University Group Diabetes Program: A Further Statistical Analysis of the Mortality Findings”. Journal of the American Medical Association 217(12):1676–87.
34. Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. “Noncoverage and Nonresponse in an Internet Survey”. Social Science Research 36:131–48.
35. Couper, Mick P. 2000. “Web Surveys: A Review of Issues and Approaches”. Public Opinion Quarterly 64(4):464–94.
36. Couper, Mick P. 2007. “Issues of Representation in eHealth Research with a Focus on Web Surveys”. American Journal of Preventive Medicine 32:S83–S89.

37. Couper, Mick P., and Michael Bosnjak. 2010. “Internet Surveys”. Handbook of Survey Research Chapter 16, P.V. Marsden and J.D. Wright editors, Bingley, UK: Emerald Group Publishing Limited.
38. Curtin, Richard, Stanley Presser, and Eleanor Singer. 2005. “Changes in Telephone Survey Nonresponse Over the Past Quarter Century”. Public Opinion Quarterly 69(1):87–98.
39. Dawber, Thomas R., Gilcin F. Meadors, and Felix E. Moore. 1951. “Epidemiological Approaches to Heart Disease: The Framingham Study”. American Journal of Public Health 41:279–86.
40. de Munnik, Daniel, David Dupuis, and Mark Illing. 2009. “Computing the Accuracy of Complex Non-Random Sampling Methods: The Case of the Bank of Canada’s Business Outlook Survey”. Bank of Canada Working Paper 2009–10, March 2009. http://www.bankofcanada.ca/wp-content/uploads/2010/02/wp09-10.pdf.
41. Dever, Jill A., Ann Rafferty, and Richard Valliant. 2008. “Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?” Survey Research Methods 2:47–62.
42. Deville, J.C., and Särndal, C.E. (1992). “Calibration estimators in survey sampling”. Journal of the American Statistical Association, 87, 376-382.
43. Deville, J.C.. 1991. “A Theory of Quota Surveys”. Survey Methodology 17:163–81.
44. Diamond, Shari S. 2000. “Reference Guide on Survey Research”. In Reference Manual on Scientific Evidence 2nd Edition. Washington, DC: Federal Judicial Center.
45. DiSogra, Charles. 2008. “River Samples: A Good Catch for Researchers?” In GfK Knowledge Networks http://www.knowledgenetworks.com/accuracy/fall-winter2008/disogra.html
46. Duffield, Nick. 2004. “Sampling for Passive Internet Measurement: A Review”. Statistical Science. 19(3):472–498.
47. Duffy, Bobby, Kate Smith, George Terhanian, and John Bremer. 2005. “Comparing Data from Online and Face-to-Face Surveys”. International Journal of Market Research 47:615–39.
48. Duncan, G. 2008. “When to Promote, and When to Avoid, a Population Perspective”. Demography 45(4):763–84.
49. Efron, Brad and Rob Tibshirani. 1993. An Introduction to the Bootstrap. CRC Press.

50. Elliott, Marc, and Amelia Haviland. 2007. “Use of a Web-Based Convenience Sample to Supplement a Probability Sample”. Survey Methodology 33(2):211–15.
51. Elliott, Michael R. 2009. “Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights”. Survey Practice. August. http://surveypractice.files.wordpress.com/2009/08/elliott.pdf.
52. Elżbieta, Getka-Wilczyńska. 2009. “Mathematical Modeling of the Internet Survey”. Warsaw School of Economics, Poland. http://www.intechopen.com/source/pdfs/8946/InTech- Mathematical_modeling_of_the_internet_survey.pdf.
53. Erikson, Robert S. and Christopher Wlezien. 2008. “Are Political Markets Really Superior to Polls as Election Predictors?” Public Opinion Quarterly 72(2):190–215.
54. Eysenbach, Gunther. 2004. “Improving the Quality of Web Surveys: The Checklist for Reporting Results from Internet E-Surveys (CHERRIES)”. Journal of Medical Internet Research 6(3):e34.
55. Fan, David P. 2011. “Representative Responses from Non-Representative Survey Samples”. Paper presented at the Midwest Association of Public Opinion Research, Chicago.
56. Field, Lucy, Rachel A. Pruchno, Jennifer Bewley, Edward P. Lemay Jr, Norman G. Levinsky. 2006. “Using Probability vs. Nonprobability Sampling to Identify Hard-to-Access Participants for Health-Related Research: Costs and Contrasts”. Journal of Aging and Health 18(4):565–83.
57. Felix-Medina, Martin H., and Steven K. Thompson. 2004. “Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations”. Journal of Official Statistics 20(1):19–38.
58. Frank, Ove. 1971. Statistical Inference in Graphs. Ph.D. thesis, Stockholm.
59. Frank, Ove. 1995. “Network Sampling and Model Fitting”. [Carrington, Peter J., John Scott, and Stanley Wasserman, eds.]. Models and Methods in Social Network Analysis. Cambridge University Press, 2005, 31–56. Cambridge Books Online. http://dx.doi.org/10.1017/CBO9780511811395.003.
60. Frankel, Martin R., and Frankel, Lester R. 1987. “Fifty Years of Survey Sampling in the United States”. Public Opinion Quarterly 51 Part 2:S127–38.
61. Fricker, Scott, and Roger Tourangeau. 2010. “Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys”. Public Opinion Quarterly 74(5):934–55.

62. Fuller, Wayne A. 1987. Measurement Error Models. New York, NY: John Wiley & Sons Inc.
63. Gile, Krista J. 2011. “Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation”. Journal of the American Statistical Association 106:135–46.
64. Gile, Krista J., and Mark S. Handcock. 2010. “Respondent-Driven Sampling: An Assessment of Current Methodology”. Sociological Methodology 40:285–327.
65. Gile, Krista J., and Mark S. Handcock. 2011. “Network Model-Assisted Inference from Respondent- Driven Sampling Data”. ArXiv Preprint.
66. Gile, Krista J., Lisa G. Johnston, and Matthew J. Salganik. 2012. “Diagnostics for Respondent- Driven Sampling”. arXiv:1209.6254. Under Review.
67. Gini, Corraco, and Luigi Galvani. 1929. “Di una Applicazione del Metodo Representative”. Annali di Statistica 6(4):1–107.
68. Gittelman, Steven H. and Elaine Trimarchi. 2009. “Consistency: The New Quality Concern”. Marketing Research Association’s Alert! Magazine, October, 49(10):19–21.
69. Gittleman, Steven H. and Elaine Trimarchi 2010. “Online Research… and All that Jazz! The Practical Adaptation of Old Tunes to Make New Music”. Online Research 2010. Amsterdam: ESOMAR.
70. Gjoka, Minas, Maciej Kurant, Carter T. Butts, and Athina Markopoilou. 2011. “Practical Recommendations on Crawling Online Social Networks”. JSAC special issue on Measurement of Internet Topologies 29(9):1872–92.
71. Glasser, Gerald J., and Gale D. Metzger. 1972. “Random-Digit Dialing as a Method of Telephone Sampling”. Journal of Marketing Research 9:59–64.
72. Goel, Sharad, and Matthew J. Salganik. 2010. “Assessing Respondent-Driven Sampling”. Proceedings of the National Academy of Science of the United States of America 107(15):6743–47.
73. Goel, Sharad, and Matthew J. Salganik. 2009. “Respondent-Driven Sampling as Markov Chain Monte Carlo”. Statistics in Medicine 28(17):2202–29.
74. Goodman, Leo A. 1961. “Snowball Sampling”. Annals of Mathematical Statistics 32:148–70.
75. Groves, Robert M. 1989. Survey Errors and Survey Costs. New York, NY: John Wiley & Sons Inc.
76. Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys”. Public Opinion Quarterly 70:646–75.

77. Groves, Robert M., and Lars Lyberg. 2010. “Total Survey Error: Past, Present, and Future”. Public Opinion Quarterly 74(5):849–79.
78. Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. “Leverage-Saliency Theory of Survey Participation: Description and an Illustration”. Public Opinion Quarterly 64:299–308.
79. Groves, Robert M., Floyd Fowler, Mick P. Couper, James Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. Survey Methodology. Wiley: New York.
80. Groves, Robert M., Stanley Presser, and Sarah Dipko. 2004. “The Role of Topic Interest in Survey Participation Decisions”. Public Opinion Quarterly 681:2–31.
81. Häder, S., and Gabler, S. 2003. “Sampling and Estimation”. In Cross-Cultural Survey Methods, Janet
82. Harkness, Fons J.R. van de Vijver, and Peter Ph. Mohler. (eds.). New York Wiley, pp. 117–34.
83. Handcock, Mark S., and Krista J. Gile. 2011. “On the Concept of Snowball Sampling”. Sociological Methodology 41(1):367–71.
84. Hansen, Morris H., William N. Hurwitz. 1943. “On the Theory of Sampling from Finite Populations”. Annals of Mathematical Statistics 14 (4):333–62.
85. Hansen, Morris H., William G. Madow, and Benjamin J. Tepping. 1983. “An Evaluation of Model Dependent and Probability-Sampling Inferences in Sample Surveys”. Journal of the American Statistical Association 78(384):776–93.
86. Harris Interactive. 2004. “Final Pre-Election Harris Polls: Still Too Close to Call but Kerry Makes Modest Gains”. The Harris Poll #87, November 2, 2004. http://www.harrisinteractive.com/harris_poll/index.asp?pid=515.
87. Harris Interactive. 2008. “Election Results Further Validate Efficacy of Harris Interactive’s Online Methodology”. Press Release from Harris Interactive, November 6, 2008.
88. Heckathorn, Douglas D. 1997. “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations”. Social Problems 44:174–99.
89. Heckman, James J. 1979. “Sample Selection Bias as a Specification Error”. Econometrica 47:153–62. http://vanpelt.sonoma.edu/users/c/cuellar/econ411/heckman.pdf.
90. Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “Beyond WEIRD: Towards a Broad- Based Behavioral Science”. Behavioral and Brain Sciences 33:111–35.

91. Intrade. 2012. http://www.intrade.com/v4/home/
92. Ipsos. 2012. “Ipsos Poll Conducted for Reuters”. Downloaded from www.ipsos- na.com/download/pr.aspx?id=11637.
93. Isaksson, Annica and Sunghee Lee. 2005. “Simple Approaches to Estimating the Variance of the Propensity Score Weighted Estimator Applied on Volunteer Panel Web Survey Data – A Comparative Study”. SRMS proceedings. 3143–3149 http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000172.pdf
94. Pasek, Josh, and Krosnick, Jon A. 2010. “Measuring Intent to Participate and Participation in the 2010 Census and Their Correlates and Trends: Comparisons of RDD Telephone and Non- Probability Sample Internet Survey Data”. Census.gov Study Series: Survey Methodology #2010- 15, Statistical Research Division, U.S. Census Bureau. https://www.census.gov/srd/papers/pdf/ssm2010-15.pdf.
95. Kalton, Graham. 1993. Sampling Rare and Elusive Populations. Department of Economic and Social Information and Policy Analysis Statistics Division, United Nations. New York.
96. Kalton, Graham. 2003. “Practical Methods for Sampling Rare and Mobile Populations”. Statistics in Transition 6:491–501.
97. Kalton, Graham. 2009. “Methods of Oversampling Rare Population in Social Surveys”. Survey Methodology December 2009, 35(2):125–41.
98. Kalton, Graham, and Dallas W. Anderson. 1986. “Sampling Rare Populations”. Journal of the Royal Statistical Society Series A, 149(1):65–82.
99. Kalton, Graham, and Ismael Flores-Cervantes. 2003. “Weighting Methods”. Journal of Official Statistics 19(2):81–97.
100. Kaplan, Charles D., Dirk Korf, Claire Sterk. 1987. “Temporal and Social Contexts of Heroin-Using Populations: An Illustration of the Snowball Sampling Technique”. The Journal Nervous and Mental Disease 175(9):566–74.
101. Kendall, Carl, Ligia Kerr, Rogerio Gondim, Guilherme Werneck, Raimunda Macena, Marta Pontes, Lisa Johnston, Keith Sabin, and Willi McFarland. 2008. “An Empirical Comparison of Respondent-Driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveillance in Men Who Have Sex with Men, Fortaleza, Brazil”. AIDS and Behavior, 12(1):97– 104.

102. Kiaer, Anders Nicolai. 1895-6. “Observations et Experiences Concernant des Denobrements Represenatifs”. Bulletin of the International Statistical Institute 9, Liv. 2:176–83.
103. Kind, Allison. 2012. Tweeting the News, Case Study: News Organizations’ Twitter Coverage of the 2011 State of the Union Address. http://www.american.edu/soc/communication/upload/Allison-Kind.pdf.
104. Kish, Leslie. 1987. Statistical Design for Research. John Wiley & Sons, New York.
105. Kish, Leslie. 1995. “The Hundred Years’ War of Survey Sampling”. Statistics in Transition 2(5):813–30.
106. Reprinted in 2003, Graham Kalton and Steven Heeringa, eds. Leslie Kish: Selected papers. New York, Wiley.
107. Kish, Leslie. 1965. “Selection Techniques for Rare Traits”. Genetics and the Epidemiology of Chronic Diseases, Public Health Service Publication No. 1163.
108. Kish, Leslie. 1965. Survey Sampling. New York: John Wiley & Sons, Inc.
109. Kleinbaum, David G., Hal Morgenstern, and Lawrence L. Kupper. 1981. “Selection Bias in Epidemiologic Studies”. American Journal of Epidemiology 113(4):452–63.
110. Klovdahl, Alden S., John J. Potterat, Donald E. Woodhouse, John B. Muth, Stephen Q. Muth, and William W. Darrow. 1994. “Social Networks and Infectious Disease: The Colorado Springs Study”. Social Science & Medicine 38(1):79–88.
111. Kogan, Steven M., Cyprian Wejnert, Yi-fu Chen, Gene H. Brody, and LaTrina M. Slater. 2011. “Respondent-Driven Sampling with Hard-to-Reach Emerging Adults: An Introduction and Case Study with Rural African Americans”. Journal of Adolescent Research 26(1):30–60.
112. Kott, Phillip S. 2006. “Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors”. Survey Methodology 32(2):133–142.
113. Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, Carolina Casas- Cordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivelore Raghunathan. 2011. “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys”. Journal of the Royal Statistical Society Series A 173(2):389–407.

114. Kruskal, William, and Frederick Mosteller. 1980. “Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939”. International Statistical Review 48:169–95.
115. Kruskal, William, and Frederick Mosteller. 1981. “Ideas of Representative Sampling”. New Directions for Methodology Of Social And Behavioral Science: Problems With Language Imprecision 3–24.
116. Lagakos, Stephen W., and Louise M. Ryan. 1985. “On the Representativeness Assumption in Prevalence Tests of Carcinogenicity”. APPLIED STATISTICS 34:54–62.
117. Lavrakas, Paul J., Charles D. Shuttles, Charlotte Steeh, and Howard Fienberg. 2007. “The State of Surveying Cell Phone Numbers in the United States–2007 Special Issue”. Public Opinion Quarterly 71(5):840–854.
118. Lee, Sunghee. 2006. “An Evaluation of Nonresponse and Coverage Errors in a Web Panel Survey”. Social Science Computer Review 2(4):460–75. http://ssc.sagepub.com/content/24/4/460.abstract.
119. Lee, Sunghee. 2006. “Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys”. Journal of Official Statistics 22(2):329–49.
120. Lee, Sunghee, and Richard Valliant. 2009. “Estimation for Volunteer Panel Web Surveys using Propensity Score Adjustment and Calibration Adjustment”. Sociological Methods and Research 37(3):319-43. http://smr.sagepub.com/content/37/3/319.full.pdf.
121. Lee, Sunghee, and Richard Valliant. 2008. “Weighting Telephone Samples Using Propensity Scores”. In Advances in Telephone Survey Methodology, edited by J. M. Lepkowski, C. Tucker, J. M. Brick, E.D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, and R. L. Sangster, 170–83. Hoboken, NJ: John Wiley & Sons, Inc.
122. Lensvelt-Mulders, Gerty J. L. M., Peter J. Lugtig, and Marianne Hubregtse. 2009. September. “Separating Selection Bias and Non-Coverage in Internet Panels Using Propensity Matching”. Survey Practice, 2 [e-journal].
123. Lesser, Virginia. “Advantages and Disadvantages of Probability and Non-Probability-Based Surveys of the Elderly and Disabled”. [online presentation]. http://ncat.oregonstate.edu/pubs/TRANSED/1081_Surveys.pdf.
124. Levy, Paul S., and Stanley Lemeshow. 2008. Sampling of Populations: Methods and Applications. 4th ed.Hoboken, NJ: John Wiley & Sons, Inc.

125. Little, Roderick J.A., and Donald Rubin. 2002. Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.
126. Little, Roderick J.A., and Sonia L. Vartivarian. 2004. “Does Weighting for Nonresponse Increase the Variance Of Survey Means?” April 2004. Working Paper 35. The University of Michigan Department of Biostatistics Working Paper Series.
127. Little, Roderick J.A. 1986. “Survey Nonresponse Adjustments for Estimates of Means”. International Statistical Review 54(2):139–57.
128. Little, Roderick J.A. 1988. “Missing-Data Adjustments in Large Surveys”. Journal of Business & Economic Statistics 63:287–296.
129. Lohr, Sharon L. 1999. Sampling: Design and Analysis. Pacific Grove, CA: Brooks/Cole Publishing Company.
130. Loosveldt, Geert, and Nathalie Sonck. 2008. “An Evaluation of the Weighting Procedures for an Online Access Panel Survey”. Survey Research Methods 2:93–105.
131. Lynn, Peter. 2004. “The Use of Substitution in Surveys”. The Survey Statistician 49:14–6.
132. Lynn, Peter, and Roger Jowell. 1996. “How Might Opinion Polls Be Improved? The Case for Probability Sampling”. Journal of the Royal Statistical Society Series A 15:21–8.
133. MacKellar, Duncan, Kathleen M. Gallagher, Teresa Finlayson, Travis Sanchez, Amy Lansky, and Patrick S. Sullivan. 2007. “Surveillance of HIV Risk and Prevention Behaviors of Men Who Have Sex With Men—A National Application of Venue-Based, Time-Space Sampling”. Public Health Reports 122(1):39–47.
134. Malhotra, Neil, and Jon A. Krosnick. 2007. “The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples”. Political Analysis 15:286–323.
135. Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. Seoul, San Francisco, London, and Washington, DC: McKinsey Global Institute. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

136. Matthews, Vince. 2008. “Probability or Nonprobability: A Survey is a Survey–or Is It?” National Agricultural Statistics Service (NASS) white paper. http://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Statistical_Aspe cts_of_Surveys/survey_is_survey.pdf
137. Mayes, Linda C., Ralph I. Horwitz, and Alvan R. Feinstein. 1988. “A Collection of 56 Topics with Contradictory Results in Case-Control Research”. International Journal of Epidemiology 17:680–5.
138. McKenzie, David. J, and Johan Mistiaen. 2009. “Surveying Migrant Households: A Comparison of Census-Based, Snowball and Intercept Point Surveys”. Journal of the Royal Statistical Society Series A, 172:339–60.
139. McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. “Birds of a Feather: Homophily in Social Networks”. Annual Review of Sociology 27:415–44.
140. Moser, Claus Adolf, and Alan Stuart. 1953. “An Experimental Study of Quota Sampling”. Journal of the Royal Statistical Society: Series A, 116:349–405.
141. Mosteller, Frederick, Herbert Hyman, Philip J. McCarthy, Eli S. Marks, David B. Truman, et al. 1949. “The Pre-Election Polls of 1948”. Report to the Committee on Analysis of Pre-Election Polls and Forecasts. Bulletin 60. New York: Social Science Research Council.
142. Muhib, Farzana B., Lillian S. Lin, Ann Stueve, Robin L. Miller, Wesley L. Ford, Wayne D. Johnson, and Philip J. Smith. 2001. “A Venue-Based Method for Sampling Hard-To-Reach Populations”. Public Health Reports 116(1):216–22.
143. Mutz, Diana. 2011. Population-Based Survey Experiments. Princeton, NJ: Princeton University Press.
144. National Research Council. 2012. Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification. Washington, D.C.: The National Academies Press.
145. New York Times. 2008. “Google Uses Searches to Track Flu’s Spread”. November 11. http://www.nytimes.com/2008/11/12/technology/internet/12flu.html?_r=2.
146. New York Times. 2012. “Which Polls Fared Best and Worse in the 2012 Presidential Race”. November 10. http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best- and-worst-in-the-2012-presidential-race/.
147. Neyman, Jerzy. 1934. “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive

Selection". Journal of the Royal Statistical Society 97:558–625.
148. O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series". In Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, Washington, DC, May 2010. www.aaai.org.
149. Office of Management and Budget (OMB). 2006. Standards and Guidelines for Statistical Surveys. http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf. (Duplicate—see U.S. Office of Management and Budget Page 7).
150. Olivier, Lex. 2011. "River Sampling Non-Probability Sampling in an Online Environment". [Web log, November 13, 2011.] Center for Information-Based Decision Making and Marketing Research. http://lexolivier.blogspot.com/2011/11/river-sampling-non-probability- sampling.html.
151. Olson, Kristen. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias". Public Opinion Quarterly 70(5):737–58.
152. Peracchi, Franco, and Finis Welch. 1995. "How Representative are Matched Cross-Sections?
153. Evidence from the Current Population Survey". JOURNAL OF ECONOMETRICS 68:153– 79.
154. Pew Research Center. 2011. "Muslim Americans: No Signs of Growth in Alienation or Support for Extremism". Survey Report. Washington, DC: Pew Research Center. http://www.people- press.org/2011/08/30/muslim-americans-no-signs-of-growth-in-alienation-or-support-for- extremism/.
155. Peytchev, Andy, Sarah Riley, Jeffrey Rosen, Joe Murphy, and Mark Lindblad. 2010. "Reduction of Nonresponse Bias in Surveys Through Case Prioritization". Survey Research Methods 4(1):21–9.
156. Pfeffermann, Danny, and C.R. Rao, editors. 2009. Sample Surveys: Inference and Analysis. Handbook of Statistics, Vol 29B. Oxford: Elsevier B.V.
157. Poynter, Ray. 2010. The Handbook of Online and Social Media Research. Chichester, United Kingdom: Wiley.
158. Presser, Stanley. 1984. "The Use of Survey Data in Basic Research in the Social Sciences". In Surveying Subjective Phenomena, vol. 2, edited by C. F. Turner and E. Martin, 93–114. New York: Russell Sage.
159. Public Works and Government Services Canada. 2008. "The Advisory Panel

on Online Public Opinion Survey Quality —Final Report". http://www.tpsgc-pwgsc.gc.ca/rop-por/rapports- reports/comiteenligne-panelonline/tdm-toc-eng.html.
160. Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data". Journal of the American Statistical Association 83:231–41.
161. Rao, John N.K. 2003. Small Area Estimation. Hoboken, NJ: John Wiley & Sons, Inc.
162. Ribeiro, Bruno, and Don Towsley. 2010. "Estimating and Sampling Graphs with Multidimensional Random Walks". In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. http://arxiv.org/pdf/1002.1751.pdf.
163. Rivers, Douglas. 2007. "Sampling for Web Surveys". White paper prepared from presentation given at the 2007 Joint Statistical Meetings, Salt Lake City, Utah, July-August. https://s3.amazonaws.com/yg-public/Scientific/Sample+Matching_JSM.pdf.
164. Rivers, Douglas, and Delia Bailey. 2009. "Inference from Matched Samples in the 2008 U.S. National Elections". Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida, May.
165. Rivers, Douglas. 2007. "Sample Matching for Web Surveys: Theory and Application". Paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, Utah, July-August.
166. Robinson, William T., Jan M.H. Risser, Shanell McGoy, Adam B. Becker, Hafeez Rehman, Mary Jefferson, Vivian Griffin, Marcia Wolverton, and Stephanie Tortu. 2006. "Recruiting Injection Drug users: A Three-Site Comparison of Results and Experiences with Respondent-Driven and Targeted Sampling Procedures". Journal of Urban Health 83(1):29–38.
167. Rosenbaum, Paul R. 2005. «Observational Study". in Everitt & Howell, eds. l Encyclopedia of Statistics in Behavioral Science. Vol. 3, 1451–1462.
168. Rosenbaum, Paul, and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". Biometrika 70:41–55.
169. Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score". Journal of the American Statistical Association 79:516–24.
170. Rothman, Kenneth J., and Sander Greenland. 1998. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott Williams & Wilkins.

171. Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. 2008. Modern Epidemiology. 3rd ed. Philadelphia: Lippincottt Williams & Wilkins.
172. Rothschild, David. 2009. “Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases”. Public Opinion Quarterly 73(5):895–916.
173. Royall, Richard. 1970. “On Finite Population Sampling Theory Under Certain Linear Regression Models”. Biometrika 57:377-87.
174. Rubin, Donald B. 2008. «For objective causal inference, design trumps analysis”. The Annals of Applied Statistics, 2, 808-840.
175. Rubin, Donald B. 1979. “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies”. Journal of the American Statistical Association 74:318–28.
176. Salganik, Matthew J. 2006. “Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling”. Journal of Urban Health 83:98–112.
177. Salganik, Matthew J., and Douglas D. Heckathorn. 2004. “Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling”. Sociological Methodology 34:193–239.
178. Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. Model-Assisted Survey Sampling. New York: Springer-Verlag, Inc.
179. Savage, Mike, and Roger Burrows. 2007. «The coming crisis of empirical sociology”. Sociology, 41, 885-899.
180. Schield, Milo. 1994. “Random Sampling Versus Representative Samples”. American Statistical Association Proceedings Of The Section On Statistical Education 107–10. Downloaded from www.StatLit.org/pdf/1994SchieldASA.pdf.
181. Schillewaert, Niels, Tom De Ruyck, and Annelies Verhaeghe. 2009. “Connected Research – How Market Research Can Get the Most Out of Semantic Web Waves”. International Journal of Market Research 51(1):11–27.
182. Schonlau, Matthias, Arthur van Soest, and Arie Kapteyn. 2007. “Are ‘Webographic’ or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?” Survey Research Methods 1:155–63.
183. Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, and Mick Couper. 2009. “Selection Bias in Web Surveys and the Use of Propensity Scores”. Sociological Methods & Research 37:291–318.

184. Schonlau, Matthias, Kinga Zapert, Lisa Payne Simon, Katherine Sanstad, Sue Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra Berry. 2004. "A Comparison Between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey". Social Science Computer Review 22:128–38.
185. Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature". Journal of Personality and Social Psychology 51:515–30.
186. Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2001. Experimental and Quasi- Experimental Designs for Generalized Causal Inference. 2nd ed. Stamford, CT: Cengage Learning/Wadsworth Publishing.
187. Shadish, William R., M. H. Clark, and Peter Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments". Journal of the American Statistical Association 103:1334–44.
188. Silver, Nate. 2012. (See New York Times 2012 above.)
189. Smith, Aaron. 2011. Trends in Cell Phone Usage and Ownership. Washington, D.C.: Pew Research Center. http://www.pewinternet.org/Presentations/2011/Apr/FTC-Debt-Collection-Workshop-Cell- Phone-Trends.aspx.
190. Smith, T. M. F. 1983. "On The Validity of Inferences From Non-Random Sample". Journal of the Royal Statistical Society Series A, 146(4):394–403.
191. Snell, Laurie, J., Peterson, Bill and Grinstead, Charles. (1998). "Chance News 7.11". Downloaded from http://www.dartmouth.edu/~chance/chance_news/recent_news/chance_news_7.11.html on August 31, 2009. Stirton Stirton
192. Snow, Rob E., John D. Hutcheson, James E. Prather. 1981. "Using Reputational Sampling to Identify Residential Clusters of Minorities Dispersed in a Large Urban Region: Hispanics in Atlanta, Georgia". Proceedings of the section on Survey Research Methods, American Statistical Association 101–6.
193. Squire, Peverill. 1988. "Why the 1936 Literary Digest Poll Failed". Public Opinion Quarterly 52:125–33.
194. Statistics Canada 2002. Statistics Canada's Quality Assurance Framework. Available from http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-586-X&CHROPG=1&lang=eng.
195. Statistics Canada. 2009. Statistics: Power from Data! Nonprobability sampling.
196. http://www.statcan.gc.ca/edu/power-pouvoir/ch13/nonprob/5214898-eng.htm.

197. Steckler, Allan, and Kenneth R. McLeroy. 2008. “The Importance of External Validity”. American Journal of Public Health 98:9–10.
198. Steiner, Peter M., Thomas D. Cook, and William R. Shadish. 2011. “On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores”. Journal of Educational and Behavioral Statistics 36:213–36.
199. Stephan, Franklin F., and Philip J. McCarthy. 1958. Sampling Opinions: An Analysis of Survey Procedure. Oxford: John Wiley & Sons, Inc.
200. Stolzenberg, Ross M., and Daniel A. Relles. 1997. “Tools for Intuition About Sample Selection Bias and its Correction”. American Sociological Review 62:494–507.
201. Stone, Mary Bishop, Joseph L. Lyon, Sara Ellis Simonsen, George L. White, and Stephen C. Alder. 2007. “An Internet-Based Method of Selecting Control Populations for Epidemiological Studies”. Practice of Epidemiology 165:109–12.
202. Strauss, Murray A. 2009. “Validity of Cross-National Research Using Unrepresentative Convenience Samples”. Survey Practice 43(3).
203. Sudman, Seymour. 1966. “Probability Sampling with Quotas”. Journal of the American Statistical Association 20:749–71.
204. Sudman, Seymour, and Brian Wansink. 2002. Building a Successful Convenience Panel. Chicago, IL: American Marketing Association.
205. Sugden, R.A. and Smith, T.M.F. 1984. “Ignorable and Informative Designs in Survey Sampling Inference”. Biometrika 71(3):495–506.
206. Taylor, Humphrey. 2007. “The Case for Publishing (Some) Online Polls”. Polling Report. Terhanian, George, and John Bremer. 2012. “A Smarter Way to Select Respondents for Surveys?” International Journal of Market Research 54(6):751–780.
207. Terhanian, George, and John Bremer. 2000. “Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research”. Research paper: Harris Interactive. http://growingupwithmedia.com/pdf/Confronting_Selection_Bias.pdf.
208. Terhanian, George, and John Bremer. 1995. “Creative Applications of Selection Bias Modeling in Market Research”. ISI paper.
209. Terhanian, George, Jonathan W. Siegel, Cary Overmeyer, John Bremer, and Humphrey Taylor. 2001. “The Record of Internet-Based Opinion Polls in

Predicting the Results of 72 Races in the November 2000 U.S. Elections". International Journal of Market Research 43(2):127–135.
210. The Telegraph. 2012. "Wisdom Index Poll Puts Labour Eight Points Ahead of Conservatives". http://www.telegraph.co.uk/comment/9307396/Wisdom-index-poll-puts-Labour-eight-points- ahead-of-Conservatives.html.
211. Thompson, Steven K. 1990. "Adaptive Cluster Sampling". Journal of the American Statistical Association 85:1050–59.
212. Thompson, Steven K. 1992. Sampling. Wiley, New York.
213. Thompson, Steven K., Linda M. Collins. 2002. "Adaptive Sampling in Research on Risk-Related Behaviors". Drug and Alcohol Dependence 68(1):S57–67.
214. Thompson, Steven K., and Ove Frank. 2000. "Model-Based Estimation with Link-Tracing Sampling Designs". Survey Methodology 26:87–98.
215. Thompson, Steven K., George A.F. Seber. 1996. Adaptive Sampling. Wiley, New York.
216. Thompson, Steven K. 2006a. "Targeted Random Walk Designs". Survey Methodology 32:11–24. Thompson, Steven K. 2006b. "Adaptive Web Sampling". Biometrics 62:1224–34.
217. Thompson, Steven K. 1990. "Adaptive Cluster Sampling". Journal of the American Statistical Association 85(412):1050–59.
218. Thompson, Steven K. 2002. Sampling. New York: John Wiley & Sons, Inc.
219. Tighe, Elizabeth, David Livert, Melissa Barnett, and Leonard Saxe. 2010. "Cross-Survey Analysis to Estimate Low-Incidence Religious Groups". Sociological Methods & Research 39:56–82.
220. Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. The Science of Web Surveys. New York: Oxford University Press.
221. Trow, Martin. 1957. Right-Wing Radicalism and Political Intolerance. Arno Press, New York, Reprinted 1980.
222. Twyman, Joe. 2008. "Getting It Right: Yougov and Online Survey Research In Britain". Journal of Elections, Public Opinion & Parties 18:343-354.
223. U.S. Census Bureau. 2011. U.S. Census Bureau Statistical Quality Standards. Washington, D.C.
224. U.S. Department of Agriculture (USDA). 2006. "The Yield Output Forecasting Program of NASS". Downloaded from http://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/yldfrcst2006.pdf.

225. U.S. Office of Management and Budget. 2006. Standards and Guidelines for Statistical Surveys. Washington, D.C. http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.
226. Valliant, Richard and Jill A. Dever. 2011. "Estimating Propensity Adjustments for Volunteer Web Surveys". Sociological Methods & Research 40(1):105–137. http://smr.sagepub.com/content/40/1/105.
227. Valliant, Richard, Alan H. Dorfman, and Richard M. Royall. 2000. Finite Population Sampling and Inference. New York: John Wiley & Sons, Inc.
228. Vavreck, Lynn and Rivers, Douglas. 2008. "The 2006 Cooperative Congressional Election Study". Journal of Elections, Public Opinion & Parties 18(4):355–66.
229. Vehovar, Vasja. 1995. "Field Substitutions in Slovene Public Opinion Survey". In Contributions to Methodology and Statistics Metodološki zvezki, 10. A. Ferligoj and A. Kramberger eds. Ljubljana: FDV, 39–66.
230. Vehovar, Vasja. 1999. "Field Substitution and Unit Nonresponse". Journal of Official Statistics 15(2):335–50.
231. Volz, Erik, and Douglas Heckathorn. 2008. "Probability Based Estimation Theory for Respondent Driven Sampling". Journal of Official Statistics 24:79–97.
232. Walker, Robert, and Raymond Pettit. 2009. ARF Foundations of Quality: Results preview. New York: The Advertising Research Foundation.
233. Watters, John K., and Patrick Biernacki. 1989. "Targeted Sampling: Options for the Study of Hidden Populations". Social Problems 36(4):416–30.
234. Wejnert, Cyprian, and Douglas D. Heckathorn. 2007. "Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research". Sociological Methods and Research 37(1):105–34.
235. Welch, Susan. 1975. "Sampling by Referral in a Dispersed Population". Public Opinion Quarterly 39:237–45.
236. Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias". Annual Review of Sociology 18:327–350.
237. Wolter, Kirk M. 2007. Introduction to Variance Estimation. New York: Springer Science+Business Media, LLC.
238. Yates, F. 1946. "A Review of Recent Statistical Developments in Sampling and Sampling Surveys". Journal of the Royal Statistical Society 109:12–43.
239. Yeager, David S., Jon A. Krosnick, LinChiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy

of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples". Public Opinion Quarterly 75:709–47.

240. Zea, Maria Cecilia. 2010. "Reaction to the Special Issue on Centralizing the Experiences of LGB People of Color in Counseling Psychology". Counseling Psychologist 38(3):425–33.
241. Zukin, Cliff, Jessica Godofsky, Carl Van Horn, Wendy Mansfield, and J. Michael Dennis. 2011. "Can a Non-Probability Sample Ever Be Useful for Representing a Population?: Comparing Probability and Non-Probability Samples of Recent College Graduates". Research presented at the 2011 Annual Conference of the American Association for Public Opinion Research. http://www.knowledgenetworks.com/ganp/docs/aapor2011/aapor11-Can-a-Non-Probability-Sample.pdf.

ОТЧЁТ РАБОЧЕЙ ГРУППЫ AAPOR О НЕСЛУЧАЙНЫХ ВЫБОРКАХ

Перевод с английского Д. Рогозина, А. Ипатовой
Научный редактор перевода А. Чуриков

Дизайн, верстка, корректура ООО «ФОМ.РУ»

Подписано в печать 14.03.2016
Формат 70x100/16
Тираж 400 экз.

Издательство ООО «Буки-Веди»
115093 г. Москва, Партийный переулок, д. 1, корп. 58.
Тел.: 8 (495) 926-63-96