Алексей Чуриков

ОСНОВЫ ПОСТРОЕНИЯ ВЫБОРКИ ДЛЯ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Фонд Общественное Мнение

Алексей Чуриков ОСНОВЫ ПОСТРОЕНИЯ ВЫБОРКИ ДЛЯ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

ФОМ

А. В. ЧУРИКОВ

ОСНОВЫ ПОСТРОЕНИЯ ВЫБОРКИ ДЛЯ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

МОСКВА
ИНСТИТУТ ФОНДА
«ОБЩЕСТВЕННОЕ МНЕНИЕ»
2020

ББК 60.5
Ч932
УДК 316:311.2

ЧУРИКОВ А. В. Основы построения выборки для социологических исследований. — М.: Институт фонда «Общественное мнение», 2020. — 240 с.— ISBN 978-5-93947-034-6.

В книге рассматриваются основные принципы построения выборки в социологических исследованиях и влияние конструкции выборки на точность получаемых результатов. Основное внимание уделяется методам конструирования выборки и практике применения этих методов в реальных исследованиях. Подробно рассматриваются широко применяемые на практике методы вероятностного отбора: систематический отбор с равной вероятностью и с вероятностью, пропорциональной размеру выборки. Техника построения статистических оценок рассматривается в минимальном объеме, необходимом для понимания принципов построения выборки.

Материал, изложенный в книге, использовался автором в течение многих лет при чтении учебного курса по методам выборочных исследований в Высшей школе экономики (НИУ ВШЭ).

При работе над книгой автор опирался на многолетний опыт работы в Фонде «Общественное Мнение». Большинство рассматриваемых в книге примеров взяты из реальных исследований.

Для студентов старших курсов социологических факультетов, а также для тех, кто интересуется теорией и практикой проведения выборочных опросов.

ISBN 978-5-93947-034-6

ОГЛАВЛЕНИЕ

К ЧИТАТЕЛЮ

Путь к изданию этой книги был долгим — целых тридцать лет.

Сначала в 1990-м при ВЦИОМе (еще том — всесоюзном) возник Фонд «Общественное Мнение» (ФОМ), куда в 1992-м пришел на работу выпускник мехмата МГУ, специалист по теории вероятностей Леша Чуриков. Тогда же — в 1992-м — ФОМ отделился и стал самостоятельным, начал проводить свои опросы населения, и Леша стал еще и спецом по построению выборок для опросов. Даже пришлось в 1996-м съездить в Мичиган и поучиться в школе знаменитого профессора Л. Киша.

И вот уже более четверти века по выборкам Алексея Чурикова каждый год выполняются более сотни общероссийских и региональных опросов, то есть в сумме на его счету получается где-то под 3 тысячи опросов. А ведь на нем лежит еще и вся обработка и анализ данных этих опросов, включая разработку программного инструментария и поддержку баз данных.

В 2002-м Алексея Владимировича пригласили в Высшую Школу Экономики, где он стал учить студентов науке и практике построения выборок для социологических опросов. При этом фабрика опросов ФОМа работала под руководством Чурикова без остановки. А он, будучи чемпионом ФОМа по дотошности и скрупулезности, окончательно перешел в своей работе на режим 24/7.

И вот однажды, году примерно в 2010-м, я сказал: «Леша, твои лекции надо издать, ведь они нужны всем, кто занимается опросами». Примерно через пару лет он ответил, что никак не получается — времени нет. Я промолчал, но довольно скоро — уже в 2015-м — воскликнул: «Профессор Чуриков! Почти все знают, что Вы — лучший! Но людям нужна твердая копия Ваших бесценных знаний! Где Ваш курс лекций, где Ваша монография, где, наконец, Ваш опус-магнум?!»

Вы бы видели его глаза в тот момент! Но правы наши респонденты: глаза боятся, а руки делают! Прошло еще каких-то пять лет, и вот я с нескрываемой радостью пишу напутствие книге Алексея Владимировича Чурикова, которой — я уверен — суждена большая и долгая жизнь на благо нашей любимой эмпирической социологии.

Александр Ослон, президент ФОМ

ОТ АВТОРА

Большинство социологических и маркетинговых исследований основаны на выборочных опросах всего населения или его отдельных категорий. Результаты таких исследований существенно зависят от того, каких людей опрашивали и как их выбирали. Вопросами правильного отбора людей и способами получения на основе этого отбора адекватных оценок занимается теория выборочных методов.

Теория выборочных методов включает множество аспектов. Цель данной книги — познакомить читателей с основными принципами построения выборки в социологических исследованиях и показать, как влияет конструкция выборки на точность получаемых результатов. Основное внимание уделяется методам конструирования выборки и практике применения этих методов в реальных исследованиях. Подробно рассматриваются широко применяемые на практике методы вероятностного отбора: систематический отбор с равной вероятностью и с вероятностью, пропорциональной размеру выборки. Техника построения статистических оценок рассматривается в минимальном объеме, необходимом для понимания принципов построения выборки.

Основой для подготовки теоретической части книги послужили работы двух классиков в области методов выборочных исследований — Лесли Киша (Leslie Kish) и Уильяма Кокрена (William G. Cochran). Книга первого «Survey Sampling» [19] и переводная книга второго «Методы выборочного исследования» [3] рекомендуются в качестве основной литературы по темам, на соответствующие разделы этих книг даются ссылки после каждой главы.

При подготовке книги автор опирался на многолетний опыт работы в Фонде «Общественное Мнение», большинство рассматриваемых в книге примеров взяты из реальных исследований. Пригодились также теоретические и практические знания, полученные автором при изучении курса по выборочным методам в Центре выборочных исследований (Survey Research Center) при Мичиганском университете (США).

Книга состоит из 11 глав. В гл. 1 даются определения и вводятся основные понятия. Главы со второй по пятую посвящены теоретическим основам выборочных методов. В гл. 2 рассматривается простая случайная выборка. Гл. 3 посвящена стратифицированной выборке. В гл. 4 описываются методы проведения систематического отбора элементов совокупности. В гл. 5 анализируется многоступенчатая кластерная выборка и разбирается метод систематического отбора кластеров с вероятностью, пропорциональной размеру. Главы с шестой по девятую посвящены вопросам практического

применения выборочных методов в реальных исследованиях. В гл. 6 подробно разбирается методика построения всероссийской выборки, описывается маршрутный метод отбора домохозяйств и способы отбора респондентов в домохозяйстве. Гл. 7 посвящена взвешиванию данных и вычислению весовых коэффициентов. В гл. 8 дается описание основных типов невероятностных выборок и применяемых в них методов отбора. В гл. 9 рассматриваются особенности построения выборки для телефонных опросов, для панельных исследований и для онлайн-опросов. В гл. 10 анализируются источники ошибок в выборочных исследованиях в рамках концепции общей ошибки исследования, характеризуются типы ошибок, возникающих на разных этапах. В гл. 11 дается краткий обзор трех тем, не затронутых в предыдущих главах. Описываются программные средства для расчета точности выборочных оценок и методы расчета, которые в них используются. Рассматривается техника минимизации изменений в выборке при повторном отборе кластеров. Дается краткое описание контролируемого отбора. В конце каждой главы приводятся ссылки на основную и дополнительную литературу для углубленного изучения материала. Разделы для необязательного чтения, которые можно пропустить без ущерба для понимания основного материала, напечатаны более мелким шрифтом.

Книга предназначена для студентов старших курсов социологических факультетов, а также для тех, кто интересуется теорией и практикой проведения выборочных опросов. Изложенный в книге материал использовался автором при чтении курса по методам выборочных исследований. Курс читался студентам-социологам Высшей школы экономики (НИУ ВШЭ) ежегодно с 2002 по 2019 гг., сначала — на пятом курсе специалитета, потом — на первом курсе магистратуры. Курс рассчитан на 60 часов.

Хочется выразить благодарность людям, которые способствовали выходу в свет данной книги. Это руководитель ФОМа Александр Ослон, который на протяжении ряда лет подталкивал меня к работе над книгой и без настойчивости которого эта она вряд ли возникла бы. Это мои первые учителя по теории выборочных методов, которые в далекие 1995–1996 гг. делились со мной своими знаниями и опытом: Елена Петренко, Михаил Косолапов, а также американские коллеги из Центра выборочных исследований (Survey Research Center) Стивен Херинга (Steven Heeringa), Джим Лепковский (Jim Lepkowski) и ушедшие из жизни Лесли Киш (Leslie Kish) и Майкл Сваффорд (Michael Swafford). Это первый декан факультета социологии ВШЭ Александр Крыштановский, тоже ушедший из жизни, который в 2002 г. пригласил меня читать курс по методам выборочных исследований. Выражаю благодарность также всем, кто работал над подготовкой и изданием этой книги.

ГЛАВА 1

ВЫБОРОЧНЫЕ ИССЛЕДОВАНИЯ, ОСНОВНЫЕ ПОНЯТИЯ

1.1. Выборочные исследования

Исследование называется *выборочным*, если в нем обследуется только часть объектов, подлежащих изучению, тогда как его результаты распространяются на все множество объектов.

Выборочное исследование служит альтернативой сплошному (или поголовному) обследованию. В социологии применяются преимущественно выборочные исследования.

Возможность судить о мнении всего общества по результатам опроса незначительной части его представителей до сих пор вызывает удивление, а иногда и недоверие. Тем не менее такая возможность основана на математической теории, которая имеет строгое математическое обоснование.

Но как и всякая математическая теория, теория выборочных методов применима при соблюдении определенных условий. Эти условия регламентируют правила отбора объектов для исследования. Прежде чем формулировать эти правила, определим основные понятия.

1.2. Совокупность, элементы совокупности

Совокупностью называют множество объектов, подлежащих изучению, а сами объекты называют *элементами совокупности*.

В социологических исследованиях элементами совокупности обычно являются люди. Однако часто объектом исследования служит не отдельный индивид, а вся его семья. В этом случае элементами совокупности являются семьи или домохозяйства. Элементами совокупности могут быть также детские сады, школы, вузы, магазины, бензоколонки, населенные пункты и т. д.

Именно элементы совокупности являются теми объектами, которые должны отбираться для выборочного исследования. Отобранные для исследования элементы называют *выборкой*.

Элементы совокупности — это элементарные единицы, которые изучаются в данном исследовании как единое целое и не подлежат расчленению на составляющие. Поэтому если исследователя интересует, например, только общий бюджет семьи, то объектом иссле-

дования является домохозяйство. Для такого исследования должна формироваться выборка домохозяйств. Но если помимо общесемейного бюджета исследователь планирует изучать структуру доходов и расходов отдельных членов домохозяйства, то элементами совокупности становятся люди, а значит, надо формировать выборку людей.

Необходимо отметить, что респонденты (т. е. те, кого опрашивают) могут не являться элементами совокупности даже в тех случаях, когда совокупность состоит из людей. Например, в исследовании детей респондентами могут быть их родители или учителя. Но при этом отбираемыми элементами должны оставаться элементы совокупности (т. е. дети), а не респонденты (родители или учителя).

На практике совокупность, которая является объектом исследования, как правило, отличается от той совокупности, из которой формируется выборка.

Совокупность, которая интересует исследователя и которую он собирается изучать, называется *изучаемой* или *целевой совокупностью* (*target population*). Совокупность, из которой производится отбор элементов, формируется выборка, называется *обследуемой совокупностью* (*sampled population* или *syrvey population*)*) [3, с. 20], но наряду с этим термином используют и другой: *основа выборки* (*sampling frame*) [3, с. 21]. Эти термины можно считать синонимами, но между ними все же есть некоторая разница. Говоря об обследуемой совокупности, обычно ограничиваются ее общим описанием, тогда как под основой выборки часто понимают структурированный массив данных, подготовленный для реализации процедуры отбора. При сложной процедуре отбора, включающей несколько этапов, для каждого этапа готовят свою основу выборки.

Вместо термина «изучаемая совокупность» могут употребляться термины «генеральная совокупность», или «юниверс» (universe), которые заимствованы из математической статистики. Л. Киш предлагал не использовать эти термины, поскольку в статистике они обозначают не совсем то, что мы вкладываем в понятие «изучаемая совокупность» [19, p. 7]. В статистике генеральная совокупность является составной частью вероятностной модели, описыва-

---

*) Термин «sampled population» можно также перевести как «выборочная совокупность». Этот перевод точнее, однако при его использовании возможна путаница, поскольку термин «выборочная совокупность» в русскоязычной литературе порой служит синонимом слова «выборка». Наряду с термином «sampled population» применяется термин «survey populations» [14, p. 70].

ющей поведение случайных величин. Генеральная совокупность, как правило, состоит из бесконечного множества элементов, а распределение случайных величин на этом бесконечном множестве описывается конкретными математическими функциями. В социологическом исследовании совокупность всегда конечна, поскольку состоит из людей, домохозяйств или конечного множества других реально существующих объектов.

Изучаемая совокупность обычно шире, чем обследуемая, при этом различия между ними могут оказаться весьма существенными. Например, во многих социологических исследованиях изучаемой совокупностью является все население страны, а определение обследуемой совокупности зависит от планируемого способа отбора респондентов. При телефонных опросах в обследуемую совокупность не попадают люди, не имеющие телефона (мобильного или стационарного). При опросах в домохозяйствах по месту жительства респондентов в обследуемую совокупность не попадают проживающие в казармах военнослужащие; люди, пребывающие в местах лишения свободы, в закрытых пансионатах, больницах, монастырях; наконец, лица без определенного места жительства. Но и те, кто имеет свой дом, могут не попасть в обследуемую совокупность по самым разным причинам. Дом может быть исключен из обследуемой совокупности из-за того, что он находится в труднодоступной местности, куда интервьюер не может добраться при помощи имеющихся у него средств или в отведенное для опроса время. Адрес дома может отсутствовать в тех справочных материалах, которыми пользуются при формировании основы выборки, и т. д.

В последующем изложении термин «совокупность» часто будет использоваться без уточнения, о какой совокупности, изучаемой или обследуемой, идет речь. В этом случае надо считать, что имеется в виду обследуемая совокупность или что различия между двумя совокупностями несущественны. Там, где различия важны, название совокупности будет даваться полностью.

1.3. Выборочные оценки

Перед выборочным исследованием всегда ставится задача по определению числовых значений параметров совокупности. Это может быть средний уровень дохода, процент людей с доходом ниже прожиточного минимума, процент людей, намеренных голосовать за ту или иную партию или политика, и т. п. Значение параметров во всей совокупности всегда будем обозначать заглавными буквами. Например, $\overline{Y}_{\text{доход}}$ — средний уровень дохода в совокупности (черта

сверху означает «среднее значение»), R — рейтинг в совокупности некоторой партии или политика.

Проведя выборочное исследование, в результате которого опрошено n респондентов, можно получить средний доход или рейтинг партии (политика) для данной выборки. Величины, посчитанные по выборке, будем всегда обозначать строчными буквами (в отличие от величин для всей совокупности). Эти величины вычисляются по следующим формулам:

$$\overline{y}_{\text{доход}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad r = \frac{m}{n},$$

где y_i — доход i-го респондента, m — число респондентов, намеренных голосовать за партию или политика, а n — размер выборки. (В принципе, по таким же формулам следовало бы вычислять значение этих параметров во всей совокупности, если бы было возможно опросить все население). Величины $\overline{y}_{\text{доход}}$ и r, полученные по выборке, называются *выборочными оценками* соответствующих параметров совокупности.

Чаще всего в качестве выборочных оценок используются средние значения величин (как в случае с доходом) или процент респондентов, удовлетворяющих определенному условию (как в случае с рейтингом). Причем процент — это тоже среднее значение, точнее, его частный случай*). Иногда используются более сложные оценки, например, оценка разности или отношения двух величин, оценка суммарного значения по всем респондентам, оценки коэффициентов регрессии или корреляции и т. п. Но для теории построения выборки разница между оценками среднего и более сложными оценками не столь принципиальна. Поэтому в дальнейшем разговор будет идти в основном о выборочных оценках среднего.

1.4. Ошибки в выборочных исследованиях

Совпадают ли выборочные оценки с истинными значениями параметров совокупности? Как правило, нет. Отклонения выборочных

---

*) Доля (или процент) людей вычисляется по общей формуле для среднего $r = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$, в которую подставляют значение $y_i = 1$ для тех респондентов, кто удовлетворяет заданному условию, и значение $y_i = 0$ для всех остальных. В итоге сумма $\sum\limits_{i=1}^{n} y_i$ равна числу m людей в выборке с заданным признаком, а формула для среднего принимает простой вид $r = m/n$.

оценок от истинных значений традиционно называют ошибками. Ошибки возникают на разных этапах проектирования и реализации исследования, каждому этапу соответствует свой тип ошибок. Приведем перечень типов ошибок с их кратким описанием.

Ошибки покрытия (coverage error) возникают, если значения параметров в изучаемой и обследуемой совокупностях различаются. Это происходит, когда имеются различия между представителями изучаемой совокупности, включенными в обследуемую совокупность и не включенными в нее.

Ошибки выборки (sampling error) возникают из-за того, что в опросе участвуют не все представители совокупности, а только их часть. Они обусловлены различиями между обследуемой совокупностью и сформированной из нее выборкой. Подробнее об ошибках выборки говорится в §1.5.

Ошибки неответов (nonresponse error) возникают из-за того, что не всех включенных в выборку респондентов удается опросить. Одни отказываются от участия в опросе или от ответов на отдельные вопросы анкеты, других не удается застать дома или дозвониться до них по телефону (в зависимости от вида опроса). Мнения тех, кого не удается опросить, часто отличаются от мнений участников опроса, а это приводит к смещениям выборочных оценок.

Ошибки корректировки (adjustment error) могут возникнуть на этапе обработки собранных данных. На этом этапе есть возможность проанализировать ошибки предыдущих этапов исследования и внести в данные корректирующие поправки. Часто корректировки делаются путем приписывания определенных весовых коэффициентов разным категориям респондентов. Например, молодым людям, которых обычно не хватает среди опрошенных, приписываются повышающие весовые коэффициенты, а пожилым женщинам, которых оказалось в избытке, — понижающие. Помимо взвешивания, существуют и другие методы корректировки. Цель корректировки — уменьшить отклонения выборочных оценок от истинных значений параметров. Но бывает, что по одним параметрам отклонения уменьшаются, а по другим — наоборот, возрастают. Так и возникают ошибки корректировки.

Ошибки измерений или наблюдений (measurement errors or errors of observation) показывают разницу между ответами респондентов на вопросы анкеты и теми понятиями или категориями, которые являются предметом исследования. В качестве простейшего примера ошибки измерений приведем ответы на вопрос о доходах респондента за прошедший месяц. Респондент может неверно понять вопрос и не отнести к доходам свои дополнительные заработки, пособия, доходы с приусадебного участка и т. п. Он мо-

жет забыть о каких-то разовых поступлениях, а может сознательно приуменьшить или завысить свой доход. Из-за этого доход респондента будет измерен с ошибкой. Если ошибки измерений по всем респондентам будут однонаправленными, все в сторону занижения или в сторону завышения доходов, то отклонения выборочной оценки от истинного среднемесячного дохода могут оказаться значительными.

Часть перечисленных ошибок, а именно ошибки покрытия и ошибки выборки, присущи только выборочным исследованиям. Остальные ошибки могут возникать как в выборочных, так и в сплошных обследованиях.

Основное внимание в данной книге уделяется решению задач, связанных с измерением и минимизацией ошибок выборки. Остальные типы ошибок рассматриваются в минимальном объеме, к ним мы вернемся в главе 10 в рамках концепции общей ошибки исследования (total survey error).

1.5. Конструкция выборки, распределение выборочных оценок, ошибка выборки

Предположим, что у нас определена обследуемая совокупность (например, это жители определенного города в возрасте 18 лет и старше), а также задана конструкция выборки, которую надо сформировать.

*Конструкция выборки**) определяется следующими параметрами:

— основой выборки — способом описания обследуемой совокупности;

— размером выборки n;

— способом отбора респондентов из совокупности;

— способом получения выборочных оценок (алгоритмом расчета).

Пусть была сформирована выборка из 1000 человек, в результате опроса которых получена выборочная оценка среднего дохода $\overline{y}_1$. Если бы в выборку попали другие 1000 человек, мы получили бы другую выборочную оценку $\overline{y}_2$, которая почти наверняка отличалась бы от оценки $\overline{y}_1$. Сформировав третью выборку, мы получили бы еще одно значение $\overline{y}_3$ и т. д. Таким образом, отбирая новых 1000 человек, мы каждый раз получали бы новую оценку.

---

*) Вместо термина «конструкция выборки» часто употребляется термин «дизайн выборки».

Возникает законный вопрос: а можно ли доверять этим разным оценкам? На самом деле ситуация не столь безрадостна, как кажется на первый взгляд. Большая часть выборочных оценок $\overline{y}_i$ мало отличаются друг от друга. При достаточно больших размерах выборки распределение оценок $\overline{y}_i$ близко к нормальному распределению и имеет вид, изображенный на рис. 1.1.

Прокомментируем этот рисунок. Предположим (чисто теоретически), что мы можем получить все возможные выборки заданной конструкции и для каждой из них посчитать выборочную оценку

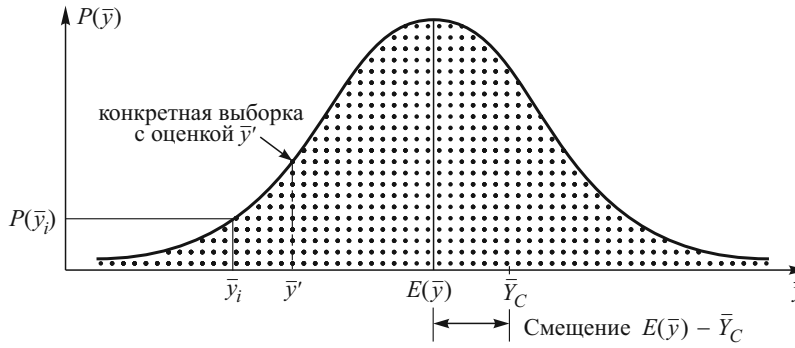

Рис. 1.1. Распределение выборочной оценки $\overline{y}$ для всего множества выборок заданной конструкции (на графике показана плотность функции распределения)

интересующего нас параметра. Поскольку совокупность конечна, число различных выборок из нее тоже конечно. Полученные значения разместим на графике следующим образом. По горизонтальной оси откладываются выборочные оценки $\overline{y}_i$ изучаемого параметра (например, среднего дохода), посчитанные по каждой отдельной выборке. По вертикальной оси для каждой такой оценки $\overline{y}_i$ фиксируется вероятность ее получения P_i. Каждую конкретную выборку можно представить в виде небольшого шарика, место размещения этого шарика на оси $\overline{y}$ определяется оценкой $\overline{y}'$ для данной конкретной выборки. Если в нескольких разных выборках получается одинаковая оценка $\overline{y}'$, то соответствующие выборкам шарики помещаются друг над другом. Чем в большем числе выборок будет получена одна и та же оценка $\overline{y}_i$, тем больше шариков будет размещено в этой точке оси $\overline{y}$ друг над другом и тем выше будет вероятность P_i получения такой оценки.

Из графика видно, что больше всего выборок будут иметь выборочную оценку $\overline{y}_i$, близкую к некоторому среднему значению, которое на графике обозначено $E(\overline{y})$. Чем дальше выборочная оцен-

ка $\overline{y}_i$ отстоит от среднего $E(\overline{y})$, тем реже (т. е. в меньшем числе выборок) она встречается.

Как же получается величина $E(\overline{y})$? Она равна усредненной выборочной оценке, посчитанной по всем возможным выборкам заданной конструкции. Для расчета $E(\overline{y})$ применяется формула

$$E(\overline{y}) = \sum_i P_i \overline{y}_i. \tag{1.1}$$

Здесь $E(\overline{y})$*) — усредненная оценка по всем возможным выборкам заданной конструкции, суммирование идет по всем оценкам $\overline{y}_i$ из этих выборок, P_i — вероятность получения оценки $\overline{y}_i$. Число различных выборочных оценок конечно, поскольку конечна сама совокупность (для бесконечной совокупности вместо знака суммы в формуле пришлось бы писать знак интеграла).

Формула (1.1) справедлива для общего случая, когда у выборок могут быть разные вероятности их получения. Если вероятность получения всех выборок одинакова (а это зависит от способа отбора респондентов), то для расчета можно использовать еще одну формулу:

$$E(\overline{y}) = \frac{1}{M} \sum_{j=1}^{M} \overline{y}_j. \tag{1.1a}$$

Здесь M — число всех возможных выборок заданной конструкции, $\overline{y}_j$ — выборочная оценка в j-й выборке; суммирование идет по всем выборкам от 1 до M. Отметим, что число слагаемых в этой и в предыдущей формулах различается. В формуле (1.1) число слагаемых равно числу различных выборочных значений $\overline{y}_i$, оно меньше M, поскольку одна и та же выборочная оценка может быть получена в нескольких выборках. В формуле (1.1а) число слагаемых равно M, а если одна и та же выборочная оценка получена в нескольких разных выборках, то при суммировании она встретится несколько раз. Напомним, что формула (1.1а) применима только для частного случая, когда вероятность получения всех выборок одинакова и равна $1/M$.

Конструкция выборки называется *несмещенной*, если среднее значение $E(\overline{y})$ выборочного распределения совпадает со средним значением $\overline{Y}_C$ оцениваемого параметра во всей обследуемой совокупности, т. е. если $\overline{Y}_C = E(\overline{y})$. Если это равенство не выполняется, то конструкция выборки называется *смещенной*, а раз-

---

*) В математике для среднего используется также термин «математическое ожидание», которое обычно обозначается символом E.

ность $\overline{Y}_C - E(\overline{y})$ называется *смещением выборки* (sampling bias). Понятие смещенности и несмещенности применимо к конструкции выборки в целом, а не к отдельной ее реализации.

Ошибкой выборки называется разница между выборочной оценкой $\overline{y}'$ конкретной выборки и средним значением $\overline{Y}_C$ в обследуемой совокупности. Она состоит из двух частей: одна — это смещение выборки $\overline{Y}_C - E(\overline{y})$, вторая — это разница между выборочной оценкой $\overline{y}'$ и средним $E(\overline{y})$ выборочного распределения.

Отметим, что значением оцениваемого параметра в обследуемой совокупности считается значение, которое получилось бы, если бы удалось опросить всех представителей обследуемой совокупности. А выборочная оценка $\overline{y}'$ — это результат опроса всех респондентов, включенных в выборку. Таким образом, ошибка выборки не включает ошибку неответов, которая возникает в ходе опроса при реализации спроектированной выборки. Ошибка выборки также не включает ошибку покрытия, поскольку измеряется только для обследуемой совокупности, и ошибку корректировки, которая может возникнуть на этапе обработки данных.

При сборе информации от респондентов, которая необходима для расчета средних оценок, применяется инструментарий (вопросник), который может быть источником ошибок измерений. Если ошибки измерений имеют место, то они присутствуют как в выборочной оценке $\overline{y}'$, так и в среднем значении $\overline{Y}_C$ для обследуемой совокупности, которое в этом случае будет отличаться от истинного значения интересующего исследователя параметра. Поэтому по ошибке выборки ничего нельзя сказать о наличии или отсутствии ошибок измерений и об их величине.

Рассмотрим подробнее одну из составляющих ошибки выборки — разницу между выборочной оценкой $\overline{y}'$ и средним $E(\overline{y})$ выборочного распределения. Можно ли по выборочной оценке $\overline{y}'$ узнать значение $E(\overline{y})$, которое в несмещенной выборке совпадает со значением $\overline{Y}_C$ в обследуемой совокупности? К сожалению, определить точное значение $E(\overline{y})$ нельзя. Выборочная оценка $\overline{y}'$ — это одна из множества оценок $\overline{y}$ всех возможных выборок данной конструкции. Неизвестно, к какой части выборочного распределения принадлежит эта конкретная выборка. Она может оказаться как из правой части распределения, так и из его левой части, а может лежать в самом центре распределения. В зависимости от этого выборочная оценка $\overline{y}'$ может оказаться больше или меньше интересующего нас значения $E(\overline{y})$, а может и совпасть с ним. Единственное, что можно сделать по выборочной оценке, — это указать интервал, в который попадает значение $E(\overline{y})$. Но даже этот интервал нельзя указать со стопроцентной уверенностью, а только с некоторой

вероятностью, пусть и весьма высокой. Например, можно указать интервал, в который попадает значение $E(\overline{y})$ с вероятностью 95% или даже 99%.

Вероятность попадания значения оцениваемого параметра в некоторый интервал называют *доверительной вероятностью*, или *уровнем доверия*, а сам интервал называется *доверительным интервалом*.

На величину доверительного интервала влияют два фактора: дисперсия распределения выборочной оценки $\overline{y}$ и требуемый уровень доверия.

Дисперсия выборочной оценки $\overline{y}$ обозначается $\mathrm{Var}(\overline{y})$*), она зависит от среднеквадратичного отклонения выборочных оценок $\overline{y}_i$ всех возможных выборок заданной конструкции от среднего выборочного значения $E(\overline{y})$. Дисперсия вычисляется по следующей формуле:

$$\mathrm{Var}(\overline{y}) = \frac{1}{M}\sum_i P_i\big[\overline{y}_i - E(\overline{y})\big]^2. \qquad (1.2)$$

Здесь $\mathrm{Var}(\overline{y})$ — дисперсия выборочной оценки $\overline{y}$, M — число всех возможных выборок заданной конструкции, суммирование идет по всем оценкам $\overline{y}_i$, которые можно получить в выборках заданной конструкции, P_i — вероятность получения оценки $\overline{y}_i$.

Если вероятности получения всех выборок одинаковы, дисперсию выборочной оценки можно вычислять по формуле

$$\mathrm{Var}(\overline{y}) = \frac{1}{M(M-1)}\sum_{j=1}^{M}\big[\overline{y}_j - E(\overline{y})\big]^2, \qquad (1.2a)$$

где M — число всех возможных выборок заданной конструкции, $\overline{y}_j$ — выборочная оценка в j-й выборке, $E(\overline{y})$ — среднее выборочного распределения, суммирование идет по всем выборкам от 1 до M. Как и в случае с формулами (1.1) и (1.1а) для вычисления $E(\overline{y})$, формула (1.2а) — это частный случай общей формулы (1.2).

От величины дисперсии $\mathrm{Var}(\overline{y})$ зависит вид функции, изображенной на рис. 1.1, а именно, ширина и крутизна образованного этой функцией «холма». Чем меньше дисперсия, тем выше и круче холм, и наоборот, чем она больше, тем холм более низкий и пологий.

---

*) От английского слова *variance* — дисперсия, отклонение.

Корень квадратный из дисперсии называется *стандартным отклонением*, или *стандартной ошибкой*, и обозначается $\mathrm{SE}(\overline{y})$:**) $\mathrm{SE}(\overline{y}) = \sqrt{\mathrm{Var}(\overline{y})}$.

Если известна выборочная оценка $\overline{y}$ и ее дисперсия $\mathrm{Var}(\overline{y})$, то значение интересующего нас параметра лежит в доверительном интервале, величина которого вычисляется по формуле

$$\Delta = t_d \mathrm{SE}(\overline{y}) = t_d \sqrt{\mathrm{Var}(\overline{y})}, \tag{1.3}$$

а границы интервала — по формуле

$$y_{1,2} = \overline{y} \pm \Delta = \overline{y} \pm t_d \mathrm{SE}(\overline{y}) = \overline{y} \pm t_d \sqrt{\mathrm{Var}(\overline{y})}. \tag{1.4}$$

Здесь Δ — величина доверительного интервала, y_1 и y_2 — его левая и правая границы, $\overline{y}$ — выборочная оценка, $\mathrm{SE}(\overline{y})$ и $\mathrm{Var}(\overline{y})$ — стандартная ошибка и дисперсия выборочного распределения, t_d — константа, зависящая от доверительной вероятности P_d, с которой гарантируется попадание оцениваемого параметра в указанный доверительный интервал. При большом размере выборки значение константы t_d можно брать из таблиц для нормального распределения, поскольку оно хорошо аппроксимирует распределение выборочной оценки $\overline{y}$ (эти значения называются квантилями). В социологии обычно используется уровень доверия $P_d = 95\%$, для него значение t_d равно 1,96. Для уровней доверия 90% и 99% значения t_d равны соответственно 1,645 и 2,576.

При небольших размерах выборки (менее 60 респондентов) распределение выборочной оценки начинает отличаться от нормального распределения. В этом случае применяются таблицы t-распределения Стьюдента, в которых константа t_d зависит не только от уровня доверия P_d, но и от размера выборки. При размере выборки n значение константы t_d надо брать для распределения с $n-1$ степенью свободы. Для вычисления доверительного интервала могут также использоваться аппроксимации другими вероятностными распределениями, подробнее об этом говорится в §§ 2.7 и 2.8.

В формулы (1.3)–(1.4) для доверительного интервала кроме выборочной оценки $\overline{y}$, которая считается по результатам опроса, и константы t_d, которая берется из справочных таблиц, входит еще дисперсия выборочной оценки $\mathrm{Var}(\overline{y})$. Откуда брать эту дисперсию? Не формировать же, в самом деле, все возможные выборки заданной конструкции, чтобы вычислить дисперсию по формуле (1.2)! Конечно, этого делать не надо. Дисперсию $\mathrm{Var}(\overline{y})$ можно

**) От английского *standard error* — стандартная ошибка.

приблизительно оценить по результатам проведенного выборочного опроса, для нее можно получить свою выборочную оценку $\mathrm{var}(\overline{y})$, аналогично тому, как для параметра совокупности $\overline{Y}$ вычисляется выборочная оценка $\overline{y}$. (Обратите внимание, что дисперсия, которая считается по всей совокупности выборок, пишется с заглавной буквы, а ее выборочная оценка пишется с маленькой буквы, т. е. тут действует то же правило, что и при обозначении параметров совокупности и их выборочных оценок.) Вопросу о том, как получить выборочную оценку дисперсии $\mathrm{var}(\overline{y})$ для разных конструкций выборки, посвящена значительная часть данной книги.

После того как получена оценка дисперсии $\mathrm{var}(\overline{y})$, можно вычислить доверительный интервал по формуле (1.3), в которую вместо дисперсии среднего $\mathrm{Var}(\overline{y})$ подставляется ее оценка. В этом интервале с доверительной вероятностью P_d лежит среднее значение $E(\overline{y})$ выборочного распределения, а для несмещенной конструкции выборки и значение $\overline{Y}_C$ оцениваемого параметра обследуемой совокупности.

От величины доверительного интервала зависит точность выборочной оценки: чем меньше доверительный интервал, тем выше точность. Поэтому под точностью оценки часто понимают величину Δ доверительного интервала. Дисперсия выборочной оценки также служит показателем точности, поскольку, зная дисперсию и умножая ее на соответствующую константу t_d, можно вычислить доверительный интервал для любой доверительной вероятности P_d.

Доверительную вероятность того, что значение оцениваемого параметра попадает в указанный интервал, можно проинтерпретировать следующим образом. Пусть, например, доверительный интервал получен для уровня доверия $P_d = 95\%$. Это означает, что если будет сформировано 100 разных выборок заданной конструкции, то в среднем для 95 из них значение параметра окажется внутри этого интервала, а для 5 выборок оно будет вне доверительного интервала. К сожалению, получив всего одну выборку, мы точно не знаем, к какой группе она принадлежит: к первым 95 или к 5 оставшимся.

Составляющая ошибки выборки, которая обусловлена отклонением выборочной оценки отдельной выборки от средней оценки $E(\overline{y})$ по всем выборкам данной конструкции и описывается доверительным интервалом, присутствует в любом выборочном исследовании (за исключением случаев, когда значения оцениваемого параметра одинаковы на всех элементах совокупности). Она зависит от случайных факторов, влияющих на отбор респондентов, ее

можно назвать случайной или *статистической погрешностью**) выборки. Вторая составляющая ошибки выборки, смещение, не является неизбежным спутником выборочной оценки. При правильной конструкции выборки смещение отсутствует.

На рис. 1.2 показано одно из возможных соотношений между общей ошибкой исследования и двумя составляющими ошибки выборки. В категорию «невыборочная ошибка» объединены все типы ошибок, кроме ошибок выборки, а именно, ошибки покрытия, неответов, корректировки и измерений.

В изображенном на рисунке варианте общая ошибка исследования суммирует все виды ошибок: к статистической погрешности добавляются выборочное смещение и невыборочные ошибки. Но

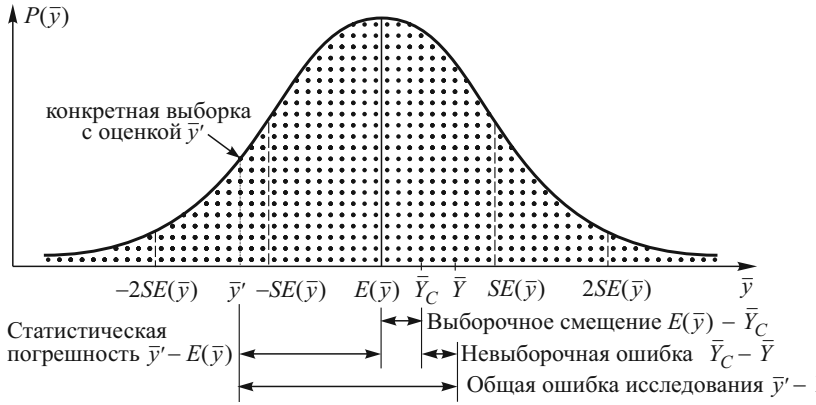

Рис. 1.2. Общая ошибка исследования

возможны и другие соотношения. Например, если бы конкретная выборка оказалась в правой части выборочного распределения, то выборочное смещение и невыборочные ошибки вычитались бы из статистической погрешности и уменьшали бы общую ошибку.

На рис. 1.2 также обозначены отклонения от среднего $E(\bar{y})$ на величины стандартной ошибки $\pm\mathrm{SE}(\bar{y})$ и удвоенной стандартной ошибки $\pm 2\mathrm{SE}(\bar{y})$. Интервал от $-2\mathrm{SE}(\bar{y})$ до $+2\mathrm{SE}(\bar{y})$ примерно соответствует 95%-му доверительному интервалу (константа $t_d = 1{,}96$ округлена до 2). Площадь участка под графиком для этого интервала равна 0,95, в него попадает 95% всех возможных

*) Использование термина «статистическая погрешность» применительно только к случайной составляющею ошибки выборки не является общепринятым, он может использоваться в более широком смысле как синоним общей ошибки исследования.

выборок заданной конструкции. Остальные 5% выборок лежат вне доверительного интервала: либо левее точки $-2\mathrm{SE}(\overline{y})$, либо правее точки $+2\mathrm{SE}(\overline{y})$. Суммарная площадь этих двух участков под графиком равна 0,05.

Дополнительно отметим различия в использовании терминов *точность* (precission) и *достоверность* (accuracy) применительно к выборочным исследованиям. Термин *точность* часто используют, говоря о величине ошибки выборки, а термин *достоверность* — характеризуя величину общей ошибки исследования [3, с. 30; 19, p. 24–25].

1.6. Вероятностные и невероятностные выборки

В §1.1 говорилось, что возможность судить о мнении всего общества по результатам опроса всего лишь незначительной части его представителей имеет строгое математическое обоснование. Но математическая теория применима только тогда, когда выборка удовлетворяет следующим двум условиям.

1. Каждый элемент совокупности должен иметь шанс (ненулевую вероятность) быть отобранным (попасть в выборку).

2. Для каждого элемента, попавшего в выборку, должна быть известна (или вычисляема) вероятность, с которой он был отобран.

Отметим, что равенство вероятностей отбора не требуется, достаточно того, что они известны или вычисляемы. Выборки, удовлетворяющие этим двум условиям, называются *случайными*, или *вероятностными*. Применительно к выборкам термины «случайная» и «вероятностная» являются синонимами (аналогичными синонимами в англоязычной литературе являются термины «random» и «probability»). Мы чаще будем использовать термин «вероятностная выборка», чтобы отличать такие выборки от случайных в обыденном смысле.

В обыденном понимании «случайным» часто называют стихийный отбор без каких-либо четких правил, когда людей для опроса произвольно отбирают на улицах или в магазинах. Для нас выборки, для которых не выполнено хотя бы одно из двух перечисленных выше условий, случайными не являются.

Сопоставим свойства вероятностных и невероятностных выборок.

Начнем со смещения. Смещение выборки возникает из-за того, что структура попадающих в выборку респондентов отличается от структуры обследуемой совокупности. Основной причиной этого служит применяемый способ отбора представителей совокупности. Они попадают в выборку с разной вероятностью, однако различия

в вероятности отбора неправильно учитываются или вовсе игнорируются исследователем. В результате в выборке оказывается излишек респондентов, вероятность отбора которых выше средней, и нехватка тех, чья вероятность попадания в выборку мала.

Если вероятность отбора респондентов никак не связана с оцениваемым параметром, то смещение выборки, скорее всего, будет небольшим (в этом случае говорят о несистематическом смещении). Если какая-то связь имеется, то смещение может оказаться значительным (тогда говорят о систематическом смещении). Систематические смещения выборки могут полностью обесценить результаты исследования.

Теория выборочного метода разрабатывалась для того, чтобы обеспечить несмещенность выборки. Применительно к вероятностным выборкам эта задача имеет решение. Если строго следовать всем теоретическим рекомендациям, то получаемая вероятностная выборка будет несмещенной, причем сразу по всем параметрам, измеряемым в исследовании. При этом процедура отбора может обеспечивать равную вероятность попадания в выборку каждому представителю совокупности (такой отбор называется равновероятностным), а может предусматривать различную вероятность отбора для разных респондентов. В последнем случае вероятность отбора должна вычисляться для каждого респондента и правильно учитываться при расчете выборочных оценок.

Для невероятностных (неслучайных) выборок невозможно правильно учесть различия в вероятности отбора респондентов, так как невозможно посчитать саму вероятность. Она зависит от разных неконтролируемых факторов, в том числе от субъективных предпочтений интервьюеров, которые участвуют в отборе. В итоге в выборке часто оказывается избыток людей, которых проще опросить, и недостаток труднодостижимых категорий респондентов. Компенсировать этот перекос выборки теми методами, которые опираются на различия в вероятностях отбора респондентов, невозможно в силу неизвестности этих вероятностей. Поэтому для невероятностных выборок не существует теории, следуя которой можно гарантировать их несмещенность. Конструкция выборки может быть смещенной по одним параметрам и несмещенной по другим, поскольку каждому измеряемому в исследовании параметру соответствует свое выборочное распределение. , Размеры смещений там, где они есть, могут заметно различаться. Например, в квотной выборке (которая не является вероятностной) смещения по квотируемым параметрам минимальны или совсем отсутствуют, тогда как по другим параметрам смещения могут быть очень большими.

В вероятностной выборке, сформированной в строгом соответствии с теорией, смещение отсутствует. Поэтому в ошибку выборки

входит только вероятностная составляющая, которая обусловлена случайным характером отбора респондентов и измеряется величиной доверительного интервала.

Можно ли говорить о доверительном интервале невероятностной выборки? В невероятностных выборках на отбор респондентов тоже влияют случайные факторы. В выборку могут попасть разные респонденты, в результате их опроса получаются разные выборочные оценки. Эти оценки имеют свое выборочное распределение со своим средним и дисперсией. Для них тоже существует доверительный интервал, в который попадает определенный процент выборочных оценок (например, 95% от всех выборок заданной конструкции). Но, в отличие от вероятностных выборок, этот интервал характеризует отклонение выборочной оценки не от значения $\overline{Y}_C$ параметра во всей обследуемой совокупности, а лишь от среднего $E(\overline{y})$ по всем невероятностным выборкам заданной конструкции, которое может существенно отличаться от $\overline{Y}_C$. Поэтому статистическая погрешность невероятностной выборки, которая задается величиной доверительного интервала, характеризует только вариативность выборочной оценки, ее возможные случайные отклонения от центра выборочного распределения.

Из-за возможных смещений невероятностной выборки, величину которых сложно оценить, значение $\overline{Y}_C$ оцениваемого параметра может не попасть в доверительный интервал с гораздо большей вероятностью, чем доверительная вероятность P_d. А если смещение велико, то $\overline{Y}_C$ может вообще не попасть в доверительный интервал для большинства выборочных оценок.

Смещенные оценки. Иногда при вычислении выборочной оценки пользуются формулами, которые также приводят к небольшим смещениям. В этом случае говорят об использовании смещенной оценки.

Смещенные оценки применяются для упрощения формул, по которым проводятся вычисления. Например, для оценивания отношения двух параметров совокупности Y/X применяют выборочную оценку y/x, которая является смещенной (подробнее о смещенности оценки y/x см. [19, section 6.6B]). Смещенные оценки рассматриваются также в § 2.10.

Смещенные оценки применяются как в вероятностных, так и в невероятностных выборках. Возникающие при этом смещения очень малы, и ими всегда можно пренебречь. Поэтому смещения этого типа представляют чисто академический интерес.

Помимо отсутствия смещений, у вероятностных выборок есть и другие преимущества перед невероятностными. Благодаря известной вероятности отбора респондентов в вероятностных выборках больше возможностей для изучения причин и оценки величины ошибок неответов, ошибок покрытия и ошибок корректировки, то-

гда как для невероятностных выборок часто бывает непросто даже отделить обследуемую совокупность от изучаемой или подсчитать долю неответивших.

Несмотря на недостатки невероятностных выборок, они широко применяются на практике. О причинах этого говорится в главе 8. Однако большинство использующих невероятностные выборки исследователей стараются максимально приблизить их к вероятностным, применяя там, где это возможно, случайные методы отбора.

1.7. Сравнение выборочных и сплошных исследований

Выборочное и сплошное исследования решают разные задачи, у каждого из них есть свои преимущества.

Выборочные исследования дешевле и требуют меньше времени. Например, всероссийский опрос населения с выборкой 1500–3000 человек можно провести меньше чем за неделю при опросе по месту жительства респондентов, а при телефонном опросе — за один-два дня; тогда как проведение всероссийской переписи населения требует не только огромных затрат, но и значительного времени на подготовку, проведение и обработку данных. Первых результатов всероссийской переписи населения 2010 года пришлось ждать почти целый год, а полностью результаты были опубликованы только через два года. Даже в таких «быстрых» сплошных обследованиях, как выборы, подведение итогов занимает несколько часов, а иногда несколько дней, тогда как результаты опроса на выходе с избирательных участков становятся известны сразу после окончания голосования*).

Основное преимущество сплошных исследований состоит в том, что они собирают информацию обо всех представителях совокупности, в них нет ни выборки, ни ошибок выборки. Предъявляя результаты сплошного исследования, не нужно указывать величину доверительного интервала и уровень доверия, как того требуют выборочные исследования. Это особенно важно для таких показателей, как общая численность населения страны или региона.

В сплошных исследованиях не возникает проблем и при анализе малых групп. Любую (сколь угодно малочисленную) часть совокуп-

---

*) Возможно, с введением автоматизированной системы для голосования ситуация с подведением итогов выборов изменится. Хотя опыт президентских выборов в США 2000 г., когда потребовался ручной пересчет бюллетеней во Флориде, позволяет в этом усомниться.

ности можно анализировать (конечно, если собрана информация обо всех ее представителях), тогда как в выборочных исследованиях препятствием для анализа малых групп является недостаточная точность выборочных оценок, связанная с малым числом представителей этих групп в выборке.

Вместе с тем, у сплошных и у выборочных исследований есть общие ошибки. Это ошибки покрытия, неответов, корректировки и измерений. Возможности влияния на величину этих ошибок различны у каждого вида исследований.

Сплошные исследования располагают широкими возможностями для уменьшения ошибок покрытия и ошибок корректировки за счет расширения границ обследуемой совокупности и за счет использования дополнительных, административных источников информации о представителях совокупности.

А вот доля неответов и особенно ошибки измерений во многом зависят от квалификации тех, кто проводит опрос. И тут выборочное исследование, которое проводится силами небольшого числа тщательно подготовленных интервьюеров, становится предпочтительнее сплошного опроса, для которого просто негде взять достаточного числа получивших необходимый практический опыт людей. Немаловажное значение имеет и тематика исследования. Трудно себе представить сплошное исследование, в котором всему населению страны задаются такие деликатные вопросы, как вопросы о пристрастии к алкоголю, к наркотикам или об интимной стороне жизни.

Поэтому выборочные исследования применяются не только для сбора социологической или политологической информации, но и для сбора статистической информации (наряду со сплошными исследованиями). В частности, часть вопросов всероссийской переписи населения 2002 года задавалась выборочно 25% респондентов (это вопросы, касавшиеся социально-экономической характеристики населения, миграции и рождаемости).

1.8. Немного истории

В 1995 году Лесли Киш (Leslie Kish) опубликовал статью [20], название которой можно перевести как «Столетняя война (или столетие войн) выборочных исследований». В статье он описывает становление выборочной теории как борьбу конкурирующих методических подходов, которые применяли исследователи в разные периоды прошлого века. Методы менялись и совершенствовались, постепенно приближаясь к достигнутому на данный момент уровню развития теории и практики.

Рождением теории выборочных исследований Киш считает публикацию в 1895 году двух статей норвежца A. N. Kiaer о репрезентативном методе (столетию с момента их выхода и посвящена статья Киша). Хотя идеи о возможности применения в исследованиях статистических вероятностных методов высказывались и раньше, в частности, еще в 1820 году. Однако первыми публикациями с обоснованием теории выборки Киш считал работы советских математиков, статьи А. А. Чупрова (начало 1900-х годов) и статью А. Г. Ковалевского «Основы теории выборочного метода» (1924). Поворотным пунктом в развитии выборочных исследований стала статья Неймана (Neyman) 1934 года, послужившая основой для дальнейшего развития теории.

После окончания Второй мировой войны теории и практики выборочных исследований стремительно развивались. Вышло сразу несколько фундаментальных работ: Йетс (Yates), 1949, Деминг (Deming), 1950, Кокрен (Cochran), 1953, Хансен, Харвитц и Медоу (Hansen, Hurwitz, Madow), 1953, Сухатми (Sukhatme), 1954, которые наряду с книгой Киша 1965 года [19] до сих пор остаются классическими учебниками по выборочным методам. В разных странах были созданы центры подготовки специалистов-выборщиков.

Развитие выборочных методов не прекращается и сейчас. Обновляется и совершенствуется инструментарий исследований, методы опросов адаптируются к постоянно меняющимся условиям современной жизни. Интернет, новые технологии, огромные массивы разнообразной информации о поведении и предпочтениях людей (big data), с одной стороны, создают конкуренцию, порождают новые альтернативные методы исследований, а с другой стороны, существенно расширяют возможности выборочных исследований.

ЛИТЕРАТУРА К ГЛАВЕ 1

Основная: [3, гл. 1], [14, sec. 2.2, 2.3, 4.1, 4.2], [19, sec. 1].

Дополнительная: [12], [14, sec. 1.2.3], [20].