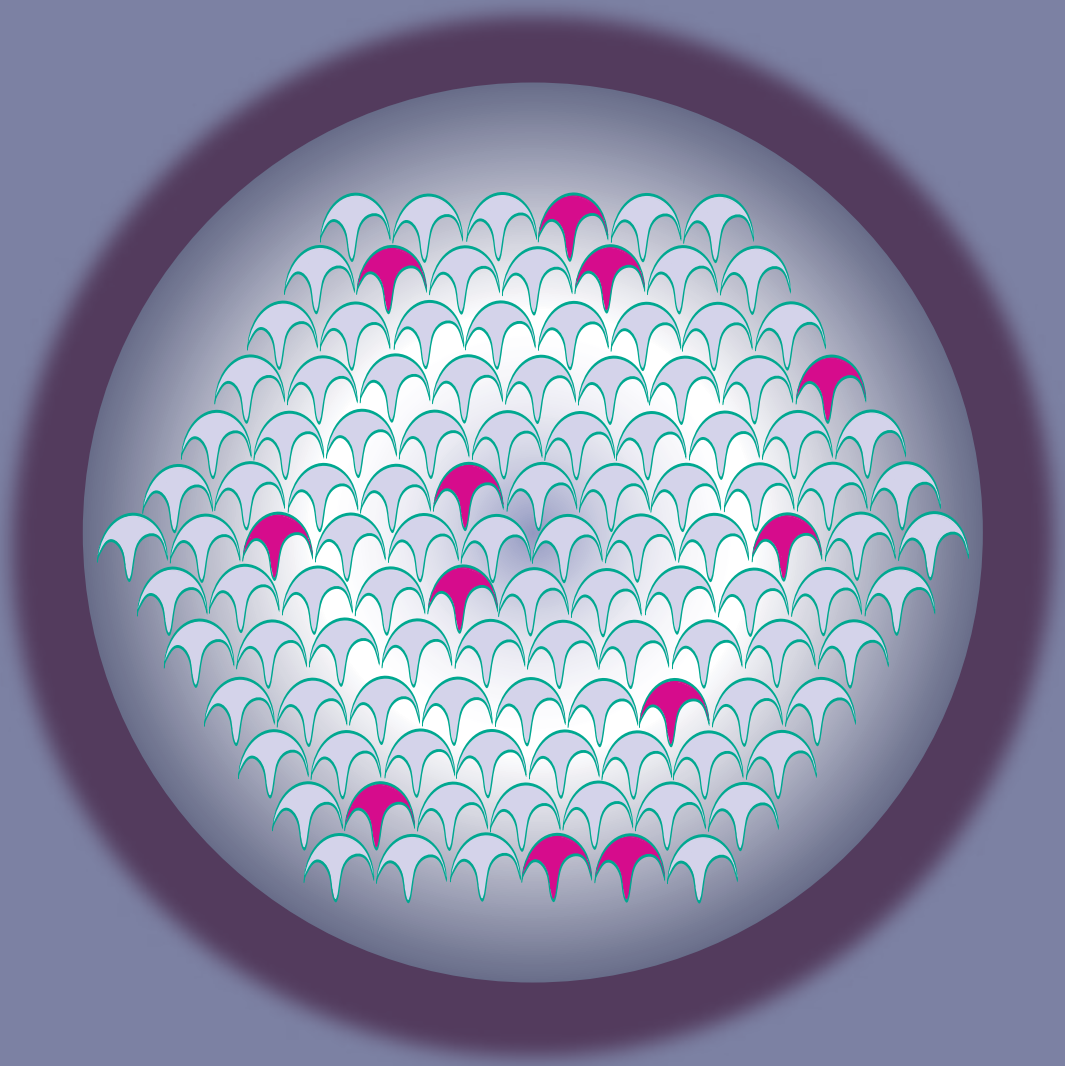
Экономические и социальные вопросы

Составление планов выборки для обследований домашних хозяйств: практические рекомендации



Организация Объединенных Наций

ST/ESA/STAT/SER.F/98

Департамент по экономическим и социальным вопросам
Статистический отдел

Методологические исследования Серия F № 98

Составление планов выборки для обследований домашних хозяйств: практические рекомендации

Организация Объединенных Наций
Нью-Йорк, 2010 год

Департамент по экономическим и социальным вопросам

Департамент по экономическим и социальным вопросам Секретариата Организации Объединенных Наций является жизненно важным звеном, обеспечивающим связь между глобальными стратегиями в экономической, социальной и экологической сферах и деятельностью на национальном уровне. Департамент ведет работу в трех основных взаимосвязанных областях: i) он собирает, вырабатывает и анализирует широкий круг экономических, социальных и экологических данных и информации, которые используются государствами — членами Организации Объединенных Наций для изучения общих проблем и критической оценки возможных стратегий; ii) он содействует проведению переговоров между государствами-членами во многих межправительственных органах с целью выработки совместных планов действий по решению существующих или возникающих глобальных проблем; и iii) он консультирует заинтересованные правительства относительно путей и средств воплощения основ политики, разработанных на конференциях и встречах на высшем уровне Организации Объединенных Наций, в виде программ на страновом уровне и по линии технической помощи оказывает содействие укреплению национального потенциала.

Примечание

Употребляемые обозначения и изложение материала в настоящей публикации не означают выражения со стороны Секретариата Организации Объединенных Наций какого бы то ни было мнения относительно правового статуса той или иной страны, территории, города или района, либо властей, или относительно делимитации их границ.

Термин «страна», используемый в настоящей публикации, также относится к территориям или районам в зависимости от контекста.

Обозначения «развитые» и «развивающиеся» страны или районы, а также «более развитые», «менее развитые» и «наименее развитые» регионы предназначены для удобства статистической работы и не означают какой-либо оценки этапа в процессе развития, достигнутого какой-либо определенной страной или районом.

Условные обозначения документов Организации Объединенных Наций состоят из прописных букв и цифр. Когда такое обозначение встречается в тексте, оно служит указанием на соответствующий документ Организации Объединенных Наций.

ST/ESA/STAT/SER.F/98

ИЗДАНИЕ ОРГАНИЗАЦИИ ОБЪЕДИНЕННЫХ НАЦИЙ

В продаже под № R.06.XVII.13

ISBN 978-92-1-461020-5

Предисловие

Основная задача публикации «Составление планов выборки для обследований домашних хозяйств: практические рекомендации» — служить руководством, которое в одном издании объединяет основные вопросы, связанные с формированием планов выборки, и к которому могли бы обращаться национальные специалисты-практики, исследователи и аналитики в области статистики, занимающиеся выборочными обследованиями в тех или иных странах. С точки зрения методологии в этом руководстве использованы обоснованные методы, закрепленные в теории статистики, которые подразумевают использование вероятностной выборки на каждом этапе процесса отбора выборочной совокупности. Правильно спланированное и надлежащим образом проведенное обследование домашних хозяйств может в короткие сроки и со сравнительно небольшими расходами дать требуемую информацию необходимого качества и точности.

Эта публикация, исходя из ее содержания, может также использоваться в качестве учебного пособия для вводных курсов по составлению планов выборки в различных обучающих учреждениях в области статистики, предлагающих курсы прикладной статистики и особенно методологии обследований.

Наряду с этим, данная публикация подготовлена в дополнение к другим публикациям по вопросам методологии выборочных обследований, которые были изданы Организацией Объединенных Наций, как, например, недавняя публикация под названием «Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой»[1], а также серия публикаций в рамках Программы обеспечения возможностей для проведения национальных обследований домашних хозяйств (ПВНОД).

В более конкретном плане задачами данного руководства являются:

a) собрать в единой публикации базовые концепции и методологически отработанные процедуры составления планов выборки, в частности, для обследований домохозяйств на национальном уровне, выделяя при этом прикладные аспекты плана выборки домохозяйств;

b) служить в качестве практического руководства для специалистов-практиков по планированию и проведению эффективных выборочных обследований домохозяйств;

c) продемонстрировать взаимозависимость между планом выборки, сбором данных, оценками, обработкой и анализом данных;

d) подчеркнуть важность контроля и сокращения *ошибок регистрации* в выборочных обследованиях домохозяйств.

Хотя читателям этого руководства и полезно иметь определенную подготовку по вопросам выборочных обследований, его смогут использовать после минимального инструктажа или вообще

1 Методологические исследования № 96 (издание Организации Объединенных Наций, в продаже под № R.05.XVII.6).

самостоятельно даже те из читателей, которые имеют лишь общие концептуальные познания в области статистики и математики. Это обусловлено тем, что одной из ключевых задач данного руководства является представление материала в применимом на практике, удобном для работы формате в отличие от акцента на теоретические аспекты выборочных обследований. При этом по мере необходимости предоставляются и теоретические основы. Чтобы без затруднений понимать излагаемый материал и применять предлагаемые методы от читателей ожидается лишь базовое знание алгебры. Соответственно, для иллюстрации тех или иных концепций, методов и способов даются многочисленные примеры.

В подготовке данного руководства участвовали целый ряд экспертов. Г-н Энтони Тернер, консультант по выборке, написал главы 3, 4 и 5 и отредактировал весь итоговый документ; г-н Ибрагим Янсанех, заместитель начальника Отдела по стоимости жизни Комиссии по международной гражданской службе, написал главы 6 и 7; и Мафион Джамбва, статистик секретариата Сообщества по вопросам развития стран юга Африки, написал главу 9.

Автором глав 1, 2 и 8, включая приложение I, являлся г-н Джереми Банда из Статистического отдела Организации Объединенных Наций, который также выполнял функции главного редактора и технического координатора проекта. Г-жа Клэр Меноцци оказала помощь в редактировании первого варианта различных глав, а г-жа Бизугенет Касса предоставила незаменимые секретарские услуги.

Проекты глав были рассмотрены на заседаниях созданной Статистическим отделом экспертной группы, которые состоялись в Нью-Йорке 3–5 декабря 2003 года. Список участников этих заседаний приводится в приложении II. Наряду с этим, данное руководство получило коллегиальную экспертную оценку со стороны доктора Альфредо Бустоса, г-жи Аны Марии Ландерос и г-на Эдуардо Риоса из Мексиканского национального института статистики, географии и информатики (INEGI), которые дали немало ценных советов.

Пол Чун
Директор
Статистический отдел Организации Объединенных Наций
Департамент по экономическим и социальным вопросам

Содержание

Стр

Глава 4
Инструментарии выборки и эталонные выборки

Глава **5**
Документирование и оценка планов выборки

Глава 6
Построение и использование весов выборки

Таблицы

Рисунки

Глава 1

Источники данных для социальной и демографической статистики

1.1. Введение

1. Во многих странах обследования домашних хозяйств входят в тройку основных источников социальной и демографической статистики. Переписи населения и жилого фонда также признаются в качестве ключевого источника социальной статистики, однако, они обычно проводятся через длительные интервалы времени порядка лет. Третьим источником являются системы административной документации. Тем не менее в большинстве стран этот источник несколько лучше разработан в области статистики здравоохранения и естественного движения населения, нежели в области социальной статистики. Обследования домашних хозяйств служат менее дорогостоящей альтернативой переписям, в том что касается обновления данных, а также более актуальной и удобной альтернативой системам административной документации. Они используются для сбора подробных и разнообразных социально-демографических данных об условиях жизни людей, их благосостоянии, осуществляемой ими деятельности, о демографических характеристиках и факторах культурного характера, влияющих на их поведение, а также о социальных и демографических изменениях. Это, однако, не мешает дополнительно использовать данные, полученные с помощью обследований домашних хозяйств, совместно с данными из других источников, таких как переписи и административные документы.

1.2. Источники данных

2. Три упомянутых выше основных источника социальных и демографических данных, если они хорошо спланированы и отработаны, а в случае систем административной документации — хорошо развиты, могут использоваться, дополняя друг друга, в рамках комплексной программы сбора и обобщения данных. Социальная и демографическая статистика абсолютно необходима для планирования и контроля программ социально-экономического развития. Статистика состава народонаселения в разбивке по возрасту и полу, включая его географическое распределение, относится к наиболее базовой категории данных, необходимых для описания населения и/или какой-либо подгруппы населения. Эти базовые характеристики дают тот контекст, в рамках которого может изучаться другая важная информация о таких социальных явлениях, как образование, инвалидность, доля работающего населения, санитарно-гигиенические условия, состояние питания, преступная виктимизация, уровни рождаемости, смертности и миграции.

1.2.1. Обследования домашних хозяйств

3. Выборочные обследования домашних хозяйств стали основным источником данных в конце 1960-1970-х годов. Они представляют собой один из наиболее гибких методов сбора данных.

В теории практически любая демографическая проблема может быть исследована с помощью обследований домашних хозяйств. Достаточно часто домашние хозяйства используются в качестве единиц выборки второго этапа в большинстве стратегий выборки с привязкой к определенным районам (см. главы 3 и 4 настоящего руководства). В выборочных обследованиях выделяется какая-то часть населения, и в рамках этого контингента проводятся наблюдения или собираются данные, затем полученные результаты экстраполируются на население в целом. Поскольку при проведении выборочных обследований на регистраторов ложится меньший объем работы, а для сбора данных отводится более длительный период времени, большинство исследуемых вопросов могут быть рассмотрены более подробно, чем в рамках переписей. Кроме того, в силу значительно меньшей потребности в сотрудниках для работы на местах, появляется возможность привлекать более квалифицированных специалистов и обучать их более интенсивно, чем при проведении переписи. На практике не все потребности той или иной страны в данных могут покрываться посредством переписей, и в силу этого фактора обследования домашних хозяйств предоставляют механизм для удовлетворения дополнительных и вновь возникающих потребностей в данных на постоянной основе. Гибкость проводимых обследований домашних хозяйств позволяет применять их в качестве отличной возможности для удовлетворения потребностей пользователей данных в той статистической информации, которая в противном случае была бы недоступной и недостаточной.

1.2.1.1. Виды обследований домашних хозяйств

4. Многие страны осуществляют программы обследований домашних хозяйств, включающие как периодические, так и специальные обследования. Рекомендуется сделать программу обследований домашних хозяйств составной частью интегрированной системы сбора статистических данных страны. В области социальной и демографической статистики частью этой системы могут стать обследования домашних хозяйств, проводимые в период между переписями.

5. Серьезной попыткой помочь развивающимся странам создать потенциал в области статистики и проведения обследований для получения требуемой социально-экономической и демографической информации по сектору домашних хозяйств стала Программа обеспечения возможностей для проведения национальных обследований домашних хозяйств (ПВНОД). Эта Программа осуществлялась в течение почти 14 лет — с 1972 по 1992 год. К моменту завершения программы в ней участвовали 50 стран. Ее главным достижением стало продвижение и принятие странами системы непрерывных многоаспектных комплексных обследований домашних хозяйств. Кроме того, эта Программа содействовала развитию потенциала в сфере выборочных обследований, особенно в африканских странах.

6. Существуют различные виды обследований домашних хозяйств, которые можно проводить для сбора данных в сфере социальной и демографической статистики. К ним относятся специализированные обследования, многоэтапные обследования, многоаспектные обследования и продольные обследования. Выбор надлежащего вида обследования зависит от целого ряда факторов, включая требования изучаемого вопроса, ресурсы и логистические соображения.

7. Специализированные обследования охватывают одну тему или вопрос, такую как использование времени или состояние питания. Обследования могут проводиться периодически или для каждого отдельного случая.

8. Многоэтапные обследования предполагают сбор статистической информации последовательными этапами, при этом каждый этап дает исходный материал для следующего этапа. Начальный этап обычно предусматривает более крупную выборку, чем последующие этапы. Его

функцией является проверка единиц выборки на предмет наличия определенных характеристик с тем, чтобы определить применимость таких единиц на последующих этапах. Эти обследования являются эффективными с точки зрения затрат способом определения контингента исследуемого населения, среди которого надлежит на дальнейших этапах собирать подробную информацию по исследуемой теме. Используя такой подход, можно из числа подходящих для изучения тем исследовать такие как, например, положение инвалидов и сирот.

9. В многоаспектных обследованиях одновременно в одном обследовании охватываются различные темы. Этот подход, как правило, является более эффективным с точки зрения затрат, чем проведение серии обследований по одной теме.

10. В продольных обследованиях сбор данных осуществляется по одним и тем же единицам выборки через определенный промежуток времени. Интервалы могут быть месячными, квартальными или годовыми. Целью проведения таких обследований является измерение изменений в некоторых характеристиках одной и той же группы населения через определенные промежутки времени. Основной проблемой этого вида обследований является высокий уровень отсева респондентов. Существует также проблема эффекта модификации выборки по заданным условиям.

1.2.1.2. Преимущества и недостатки обследований домашних хозяйств по сравнению с переписями населения

11. Хотя обследования домашних хозяйств и не являются столь же дорогостоящими, как переписи, с ними, тем не менее, могут быть связаны значительные расходы, если требуется получение показателей по отдельным административным единицам довольно низкого уровня, таким как провинции или районы. В отличие от переписи, когда сбор данных осуществляется по миллионам домашних хозяйств, выборочное обследование из-за финансовых соображений, как правило, лимитировано выборкой в несколько тысяч домохозяйств, что серьезно ограничивает возможности получения надежных данных для небольших территорий. Взаимозависимость между размером выборки и надежностью данных для небольших территорий и административных единиц рассматривается в последующих главах.

12. Обследования домашних хозяйств имеют следующие преимущества по сравнению с переписями:

a) Как указывалось выше, общие расходы на проведение обследования обычно ниже по сравнению с переписью, поскольку последняя требует больших трудовых, финансовых, логистических и материальных ресурсов. Надлежащим образом сформированная и обследованная вероятностная выборка даст точные и надежные результаты, которые могут служить основой для того, чтобы сделать определенные выводы относительно населения в целом. Вследствие этого, для некоторых показателей типа общего коэффициента рождаемости нет насущной необходимости в проведении переписи.

b) Как правило, выборочные обследования дают статистическую информацию более высокого качества, поскольку, как указывалось выше, они расширяют возможности привлечения более квалифицированных и подготовленных регистраторов. Также обследования лучше поддаются контролю, поскольку контролеры обычно хорошо подготовлены, и соотношение контролер/регистратор может достигать 1 к 4. Наряду с этим существует возможность использовать в обследованиях техническое оборудование более высокого качества для проведения физических измерений, когда такие измерения необходимы. В случае переписи качество данных иногда страдает из-за масштабности всего мероприятия. Это делает перепись уязвимой для упущений или отсутствия внимания к контролю

качества на различных этапах, что приводит к высокому уровню ошибок регистрации данных.

c) Выборочное обследование по сравнению с переписью характеризуется большей глубиной и гибкостью, а также бóльшим числом позиций вопросника. В рамках переписи может отсутствовать возможность сбора информации по более специальным вопросам, поскольку для проведения такого исследования потребуется привлечь непомерно большое число специалистов или оборудования. Вряд ли возможным окажется, например, взвешивание продуктов питания и проведение других измерений в рамках исследования адекватности питания населения. Аналогичным образом нецелесообразно будет подвергать каждого жителя страны медицинскому обследованию для определения, например, частоты случаев инфицирования ВИЧ/СПИДом. Более того, существует возможность добавлять в выборочные обследования домашних хозяйств такие вопросы, которые являются слишком сложными для переписи.

13. Выборочные обследования являются более подходящими для сбора данных на уровне стран и сравнительно крупных административно-географических единиц по таким требующим углубленного исследования темам, как многофакторные аспекты инвалидности, расходы домашних хозяйств, деятельность работающего населения и преступная виктимизация. В этом они отличаются от переписей, которые служат источником собираемой в процессе их проведения сравнительно общей информации, охватывающей небольшие административные единицы.

14. Как правило, к сильным сторонам статистических мероприятий в рамках обследований домашних хозяйств можно отнести достаточную гибкость инструментов сбора данных для того, чтобы включить большее число вопросов по разнообразной тематике, а также возможность оценки параметров, сопоставимых по количеству с теми, измерение которых ведется с помощью переписей населения и жилого фонда.

1.2.2. Переписи населения и жилого фонда

15. Перепись населения, ниже именуемая «перепись», включает весь процесс сбора, обобщения, оценки и распространения демографических, социальных и других данных, охватывающих на определенный момент времени население всей страны или четко ограниченной части или частей данной страны. Перепись является основным источником социальной статистики, обладая при этом очевидным преимуществом в плане предоставления надежных данных — иными словами, данных, не подверженных ошибкам выборки, — по малым административно-территориальным единицам. Перепись служит идеальным средством предоставления информации о численности, составе и географическом распределении населения в дополнение к социально-экономическим и демографическим характеристикам. Как правило, в рамках переписи ведется сбор информации по каждому отдельному лицу в домашнем хозяйстве и по каждому комплексу жилых помещений — обычно на всей территории страны или в каких-либо четко определенных ее частях.

1.2.2.1. Основные характеристики традиционной переписи населения и жилого фонда

16. Традиционные переписи населения и жилого фонда характеризуются следующими особенностями:

a) ведется отдельная регистрация каждого лица и каждого комплекса жилых помещений, а также их характеристик;

b) целью является охват всего населения, проживающего на четко определенной территории. В перепись включается каждое присутствующее и/или постоянно проживающее лицо в зависимости от используемой формы подсчета населения — фактической или юридической. В отсутствие всеобъемлющих реестров населения или административных реестров переписи являются единственным источником статистики по небольшим территориям;

c) как правило, регистрация по всей стране проводится по возможности одновременно. Все люди и жилые помещения регистрируются по состоянию на один и тот же учетный период;

d) переписи обычно проводятся с определенной периодичностью. В большинстве стран переписи проводятся каждые 5 лет, в некоторых странах — каждые лет. Это обеспечивает наличие сопоставимой информации, получаемой через фиксированные интервалы времени.

1.2.2.2. Использование результатов переписи

17. Ниже приводятся сведения об использовании результатов переписи:

a) переписи дают информацию о численности, составе и географическом распределении населения вместе с его демографическими и социальными характеристиками;

b) переписи служат источником статистики по небольшим территориям;

c) счетные участки переписи являются основным источником инструментария выборки для обследований домашних хозяйств. Собранные в ходе переписи данные зачастую используются в качестве вспомогательной информации в целях стратификации выборки и повышения качества оценок при обследовании домашних хозяйств.

1.2.2.3. Основные недостатки переписей

18. В силу своего уникального географического охвата перепись обычно служит основным источником базовых данных по характеристикам народонаселения. В связи с этим, однако, в рамках переписи нецелесообразно в особых подробностях охватывать многие темы. Перепись не является самым идеальным источником детальной информации, например, по экономической активности. Такая информация требует получения ответов на подробные вопросы и использования наводящих вопросов.

19. В силу того, что в рамках переписи опрос часто проводится с привлечением третьих лиц в качестве респондентов, получаемая информация часто не отличается точностью в плане тех параметров, которые могут быть известны только самому регистрируемому лицу, а именно — род деятельности, отработанное время, доход и т. д.

20. За последние несколько десятилетий переписи населения были проведены во многих странах. Например, в ходе раунда переписей 2000 года (1995–2004 годы) их провели около 184 стран и территорий.

1.2.3. Административные документы

21. Многие виды данных социальной статистики собираются из различных административных документов в соответствии с результатами ряда административных процедур. К таким примерам

можно отнести получаемую статистику: о здравоохранении — на основе отчетности больниц, о занятости — на основе данных от бирж труда, о естественном движения населения — на основе данных от учреждений регистрации актов гражданского состояния и о положении в сфере образования — на основе данных о зачислении в учебные заведения из отчетности министерств образования. Надежность статистических данных, получаемых из административных документов, зависит от полноты такой документации и единообразия определений и понятий.

22. Хотя административная документация может стать весьма эффективным с точки зрения расходов источником данных, в большинстве развивающихся стран такие системы недостаточно разработаны. Это означает, что, как правило, такие данные не отличаются точностью. Даже если для административных целей процедуры составления документации характеризуются последовательностью, сбор статистических данных в большинстве случаев и для большинства организаций является второстепенной задачей, в результате чего страдает качество таких данных. Обычно не принимаются во внимание или не соблюдаются обязательные к выполнению статистические требования, такие как стандартизация понятий и определений, своевременность и полнота охвата.

23. В большинстве стран информация из административных документов зачастую имеет ограниченный по своему содержанию характер, поскольку такие документы в большей степени применяются для юридических или административных целей. Примерами совершенствования административных систем во многих странах могут служить учреждения регистрации актов гражданского состояния. Тем не менее далеко не все страны преуспели в этой области. Страны с полномасштабными системами регистрации естественного движения населения способны периодически издавать отчетность по актам гражданского состояния, такую как число живорожденных детей в разбивке по полу, дате и месту рождения; число смертей в разбивке по возрасту, полу, месту и причине смерти; браки и разводы и т. д.

24. В рамках реестров населения ведутся базы данных с историей жизни каждого человека и каждого домашнего хозяйства в стране. Этот реестр постоянно обновляется при изменении данных о том или ином лице и/или домохозяйстве. Если такие реестры объединены с другими аналогичными системами социальной сферы, они могут служить хорошим источником информации. Подобные системы разработаны в таких странах, как Дания, Норвегия, Нидерланды, Германия и Швеция. В большинстве этих стран переписи базируются на таких системах регистрации.

25. Во многих развивающихся странах система административной документации по различным социальным программам разработана слабо, хотя она и весьма привлекательна и может служить экономически эффективным источником данных. Административные документы зачастую ограничены по своему содержанию и обычно не столь широко применимы, как данные обследований домашних хозяйств с точки зрения детализации концепций или рассматриваемых вопросов. В этом случае их использование в качестве дополнения к другим источникам информации становится весьма сложной задачей в силу отсутствия стандартизованных концепций, систем классификации вкупе с избирательным или недостаточным охватом.

1.2.4. Взаимодополняемость трех указанных источников данных

26. В данной главе указаны различные способы, с помощью которых можно согласованно использовать результаты переписей, обследований и систем административной документации. В настоящем разделе более подробно рассматривается проблема взаимодополняющего объединения информации из различных источников данных. Интерес к этой проблеме обусловлен необходимостью таких мер, как ограничение расходов на проведение переписей и обследований, снижение нагрузки на респондентов, предоставление данных по административным единицам

более низкого уровня, которые могут и не быть охвачены данными по результатам обследования, и максимально возможное использование имеющихся в стране данных.

27. Поскольку переписи невозможно проводить часто, обследования домашних хозяйств создают основу для обновления некоторой информации переписей, особенно на уровнях страны или ее крупных административных единиц. В большинстве случаев в рамках переписей исследуются лишь сравнительно общие темы, и количество вопросов обычно ограничено. Следовательно, информация переписей может быть дополнена более подробной информацией по комплексным проблемам, получаемой из обследований домашних хозяйств, пользуясь преимуществом их небольшого масштаба и потенциальной гибкости.

28. Во многих случаях переписи и обследования домашних хозяйств носят взаимодополняющий характер. Сбор информации в ходе переписи по дополнительным темам из выборки домашних хозяйств является экономически эффективным способом расширения масштабов переписи для удовлетворения растущих потребностей в социальной статистике. Использование выборочных методов и способов позволяет получать в срочном порядке необходимые данные с приемлемой точностью, когда ограниченные сроки и расходы делают практически невозможным получение таких данных посредством всеобщей переписи.

29. Переписи также создают инструментарий выборки, статистическую инфраструктуру, потенциал статистической деятельности и исходные статистические данные, необходимые для проведения обследований домашних хозяйств. Довольно распространено формирование выборки домашних хозяйств в рамках переписи в целях сбора информации по более сложным темам, таким как показатели инвалидности, материнской смертности, экономической активности и рождаемости.

30. Переписи служат основой для обследований домашних хозяйств, предоставляя инструментарий выборки: переписи дают полные списки всех административных единиц, таких как счетные участки, обычно используемые в качестве единиц первого этапа в процессе отбора домашних хозяйств для обследования. Более того, определенная вспомогательная информация, получаемая в ходе переписи, может использоваться для эффективного планирования обследований. К этому следует добавить, что вспомогательная информация по результатам переписи может использоваться для уточнения показателей выборки посредством регрессии и оценки соотношений, повышая таким образом точность показателей обследования.

31. Для обеспечения интеграции источников данных необходимо четко определить регистрационные единицы и выбрать сопоставимые административно-территориальные единицы при сборе и представлении статистической отчетности из различных источников. Кроме того, требуется выбрать общие определения, понятия и классификации по различным источникам данных, включая административную документацию.

32. Данные из результатов обследований домашних хозяйств могут также использоваться для проверки охвата и содержания переписи в целях определения размера и области ошибок. В ходе раунда переписей 2000 года в Замбии и Камбодже для этой цели использовались, например, проведенные после переписи обследования, которые позволили оценить ошибки охвата. Аналогичным образом данные переписей могут использоваться для оценки некоторых результатов обследования.

33. Получение оценок для небольших районов, чему уделяется столь большое внимание в связи с растущей потребностью в надежных оценках для таких территорий, — это именно та область, в которой используются данные обследований и административных документов для получения

параллельных оценок. Традиционные прямые оценки по конкретным районам не обеспечивают надлежащей точности, поскольку размеры выборки на небольших территориях редко являются достаточно крупными. Проведение оценок для небольших районов основывается на целом спектре статистических методов, используемых для оценивания на тех или иных территориях в случаях, когда традиционные результаты обследования для таких территорий ненадежны или не могут быть рассчитаны. Эти методы включают модели, обеспечивающие связь с соответствующими небольшими территориями посредством дополнительных или вспомогательных данных, таких как данные последней переписи населения. Основная идея этих процедур для небольших территорий состоит, таким образом, в заимствовании и объединении сравнительных преимуществ различных источников данных в попытке получить более точные и надежные оценки.

34. В странах с хорошо развитыми системами регистрации актов гражданского состояния данные переписей и обследований могут успешно использоваться совместно с данными из административных документов. Например, в ходе переписи населения 1990 года в Сингапуре регистраторы имели заранее заполненные формуляры с базовой информацией из административных документов по каждому члену домашнего хозяйства. Этот подход сократил время опроса и расходы на проведение регистрации. Поскольку основанная на реестрах перепись дает только общую численность населения и его базовые характеристики, подробные социально-экономические характеристики собираются на основе выборки.

35. Данные из административных документов могут использоваться для проверки и оценки результатов обследований и переписей. Например, в странах с полномасштабной системой регистрации естественного движения населения данные по рождаемости и смертности, полученные из переписи, могут сверяться с данными системы регистрации.

1.2.5. Заключительные замечания

36. В заключение следует отметить, что обследования домашних хозяйств, переписи и административные источники должны рассматриваться как взаимодополняющие источники. Это означает, что при планировании переписей и обследований при любой возможности следует использовать общие понятия и определения. Необходимо также периодически проверять административные процедуры для обеспечения использования общих понятий и определений.

37. Для адекватного удовлетворения общих потребностей в социально-демографической статистике программа обследований домашних хозяйств должна входить в качестве составной части в интегрированную систему сбора статистических данных страны, включающую также переписи и административные документы.

Справочная литература и дополнительные источники информации

Ambler, R. and others (2001). Combining unemployment benefits data and LFS to estimate ILO unemployment for small areas: an application of the modified Fay-Herriot method. Invited paper. International Statistical Institute session, Seoul.

Banda, J. (2003). Current status of social statistics: an overview of issues and concerns. Presented at the Expert Group Meeting on Setting the Scope of Social Statistics organized by the United Nations Statistics Division in collaboration with the Siena Group on Social Statistics, New York, 6–9 May 2003.

Bee-Geok, L., and K. Eng-Chuan (2001). ESA/STAT/AC.88/05 7 April. Combining survey and administrative data for Singapore's census of population 2000. Invited paper, International Statistical Institute session, Seoul.

Kiregyera, B. (1999). *Sample Surveys: With Special Reference to Africa*. Kampala: PHIDAM Enterprises.

Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology* (Statistics Canada, Ottawa), vol. 25, No.2, pp. 175-186.

Singh, R. and N. Mangat (1996). *Elements of Survey Sampling.* Boston, Massachusetts: Kluwer Academic Publishers.

Statistics Canada (2003). *Survey Methods and Practices.* Ottawa.

United Nations (1982). *National Household Survey Capability Programme*: Non-sampling errors in household surveys: Sources, Assessment and Control. Preliminary version. DP/UN/INT-81-041/2., New York: United Nations Department of Technical Cooperation for Development and Statistical Office.

Организация Объединенных Наций (1984). *Руководство по обследованию домашних хозяйств* (Переработанное издание). Методологические исследования, № 31; в продаже под № R.83. XVIII.13.

_____ (1998). *Принципы и рекомендации по проведению переписей населения и жилого фонда*, Первое пересмотренное издание. Статистические документы, № 67/Rev.1. В продаже под № R.98. XVII.8.

_____(2002). *Принципы и рекомендации для системы статистики естественного движения населения,* Второе пересмотренное издание. В продаже под № R.01.XVII.10. ST/ESA/STAT/SER.M/19/Rev.2.

United Nations (2002). Technical report on collection of economic characteristics in population censuses. New York and Geneva: Statistics Division, Department of Social and Economic Affairs, and Bureau of Statistics, International Labour Office. ST/ESA/STAT/119.

Whitfold, D. and J. Banda, (2001), Post enumeration surveys (PES's): are they worth it? Presented at the Symposium on Global Review of Round of Population and Housing Censuses: Mid–decade Assessment and Future Prospects, New York, 7–10 August organized by the Statistics Division, Department of Economic and Social Affairs, United Nations Secretariat. Symposium 2001/10.

Глава 2

Планирование и проведение обследований

1. Хотя основной упор в данном руководстве делается на тех аспектах обследований домашних хозяйств, которые связаны с выборкой, необходимо дать общий обзор планирования, организации и проведения обследований, для того чтобы главы и разделы по вопросам выборки воспринимались в соответствующем контексте. Существует множество учебников, справочников и руководств, в которых весьма детально рассматриваются вопросы планирования и проведения обследований домашних хозяйств и к которым читателей настоятельно просят обращаться за дополнительной информацией. Тем не менее многие основные моменты рассмотрены и кратко описаны в данной главе, включая важнейшие аспекты планирования и проведения обследований, кроме составления планов и формирования выборки, которые рассматриваются в главах 3, 4 и в приложении I.

2.1. Планирование обследований

2. Для того чтобы обследование дало желаемые результаты, необходимо уделять особое внимание подготовительным мероприятиям, предшествующим работе на местах. В этой связи все обследования для достижения успеха требуют тщательной и продуманной подготовки. Однако объем планирования варьируется в зависимости от вида обследования, требуемых материалов и информации. Поскольку разработка надлежащего плана обследования требует достаточного количества времени и ресурсов (см. рисунок 2.1), то применительно к комплексным обследованиям нередким является использование двухгодичного цикла планирования (для более подробной информации по вопросам планирования обследований, см. издание Организации Объединенных Наций, 1984 год).

Рисунок 2.1 может служить иллюстрацией расписания проведения обследования.

2.1.1. Цели обследования

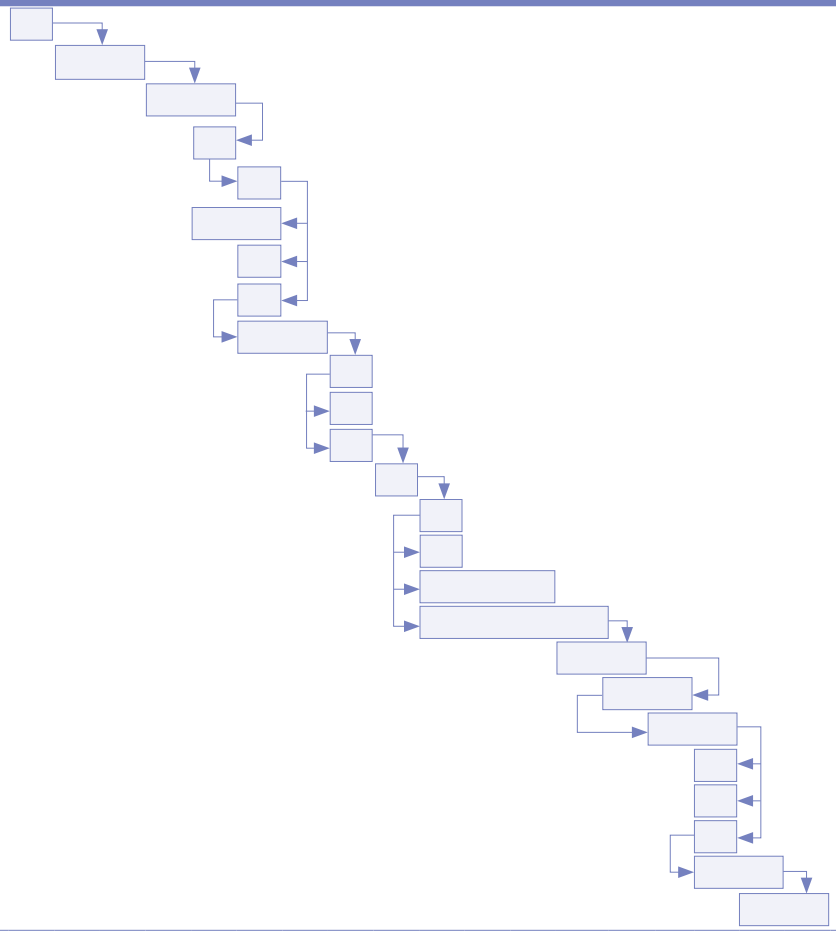
3. Обязательным условием является четкая постановка целей обследования с самого начала осуществления проекта. Должно быть четко сформулировано задание на требуемую статистическую информацию с ясным описанием населения и географического района, охватываемого обследованием. На этом этапе необходимо также указать, как будут использоваться полученные результаты. Выделенный на обследование бюджет должен служить руководством для статистиков обследования в определении задач. Четкое понимание бюджетных ограничений облегчит успешное планирование и проведение обследования.

4. В некоторых случаях цели обследования не устанавливаются явным образом. Например, проводящая обследование организация может быть привлечена к проведению исследования деятельности неформального сектора экономики. Если цель ясно не обозначена, то перед стати-

Рисунок 2.1

Расписание мероприятий в рамках обследования домашних хозяйств в стране X

№	Задача	Срок
1	Встреча с заинтересованными сторонами	1 месяц
2	Подготовительные мероприятия	3 месяца
3	Разработка первичного вопросника	2 месяца
4	Рассылка вопросника членам комитета пользователей	1 месяц
5	Включение в вопросник согласованных вопросов	1 месяц
6	Разработка проекта пособий для регистраторов и контролеров	2 месяца
7	Печатание вопросников и пособий	1 месяц
8	Составление плана предварительного испытания	1 месяц
9	Подготовка регистраторов и контролеров	2 месяца
10	Проведение предварительного испытания	1 месяц
11	Внесение изменений в вопросник (при необходимости)	1 месяц
12	Внесение изменений в пособия (при необходимости)	1 месяц
13	Создание плана выборки	1 месяц
14	Разработка и тестирование программ(ы) ввода данных	1 месяц
15	Разработка и тестирование программ(ы) очистки данных	1 месяц
16	Сбор данных	3 месяца
17	Ввод данных	4 месяца
18	Разработка и тестирование плана табулирования данных	2 месяца
19	Очистка данных (например, проверка диапазона и соответствия)	2 месяца
20	Распределение написания глав между авторами	2 месяца
21	Проведение обработки данных в соответствии с планом	1 месяц
22	Расчет ошибок выборки среди других полученных оценок	1 месяц
23	Распределение таблиц между авторами	1 месяц
24	Анализ данных и подготовка отчета	2 месяца
25	Обобщение данных и подготовка таблиц итогового отчета	2 месяца



стиком или руководителем обследования стоит задача дать определение неформального сектора с функциональной точки зрения в целях обследования с детальным выделением конкретных видов экономической деятельности, которые наиболее точно отражают требования организации-заказчика. Следует отметить, что, если цели того или иного обследования представляются неопределенными и расплывчатыми, в нем будет присутствовать высокая доля ошибок регистрации данных обследования.

5. Очень важно, чтобы к определению цели обследования, а также его масштаба и охвата привлекались заинтересованные стороны, т. е. различные пользователи и поставщики статистических данных. Проведение консультаций помогает прийти к общему мнению или компромиссу по поводу того, какие данные требуются, в какой форме необходимы эти данные, каковы уровни разукрупнения полученных данных, стратегии распространения результатов и показатели периодичности сбора данных.

6. Некоторые из обследований, проведенных соответствующими организациями, имеют четко обозначенные цели. Например, Замбийское экспериментальное обследование рабочей силы 1983 года имело следующие цели:

a) собрать информацию по численности и составу работающего на данный момент населения в формальном секторе экономики;

b) оценить спрос и предложение рабочей силы;

c) служить основой для составления прогнозов состояния рабочей силы по ряду профессий;

d) оказать помощь в планировании повышения потенциала системы образования в тех областях, которые играют ключевую роль для экономического развития.

7. Следует отметить, что наличие четко заявленных целей является первым шагом в определении тех вопросов обследования, по которым необходимо получить статистические ответы.

2.1.2. Статистическая совокупность обследования

8. При планировании обследования необходимо определить охватываемые им географические районы и исследуемое население. Например, при обследовании доходов и расходов домохозяйств могут охватываться городские районы и, возможно, исключаться сельские районы.

9. При определении статистической совокупности необходимо совершенно точно выявить группу населения, из которой будет делаться выборка. В обследованиях доходов и расходов домохозяйств совокупностью единиц первого этапа будут являться счетные участки (т. е. административно-территориальные единицы, распределенные, например, по всей территории страны), а единицами второго этапа будут домашние хозяйства в отобранных счетных участках (в главах 3 и 4 кластеры рассматриваются более подробно, кроме того, определение кластера дается в приложении I).

10. Нужно отметить, однако, что на практике численность обследуемого населения несколько меньшая, чем все формирующее совокупность население. Ограничение обследуемого населения является обычным подходом по целому ряду причин. Из некоторых обследований могут исключаться определенные домохозяйства военнослужащих, проживающих в казармах. В обследованиях рабочей силы дети младше определенного возраста могут указываться, как члены обследуемых домохозяйств, но при этом они не будут относиться к рабочей силе.

11. Важно подчеркнуть, что когда фактическое население отличается от обследуемого населения, результаты относятся к той определенной части населения, из которого делалась выборка. Как указано в главе 4, на каждом этапе отбора должны формироваться всеобъемлющие и взаимоисключающие инструментарии выборки.

2.1.3. Подлежащая сбору информация

12. Из списка вопросов, требующих статистических ответов, может быть получен перечень пунктов, которые могут предоставить фактическую информацию, значимую для исследуемой проблемы. Важно всегда иметь в виду, что некоторые требуемые данные могут быть получены из уже имеющихся источников. При составлении перечня пунктов необходимо предусмотреть включение в него дополнительных пунктов, которые будут взаимосвязаны с основными пунктами. Например, в рамках обследования занятости и заработков может собираться дополнительная информация о возрасте, половой принадлежности и образовании. Такая информация улучшит понимание смежных проблем и, таким образом, позволит провести более глубокий анализ.

13. Можно также добавить, что на этапе планирования обследования должен составляться план табулирования данных. Бланки предлагаемых таблиц должны быть разосланы для получения замечаний и предложений по их улучшению.

2.1.4. Бюджет обследования

14. Бюджет обследования отражает финансовые потребности в связи с проведением данного обследования. Бюджет необходим для поддержки и обеспечения проведения намеченного обследования, а также для составления графика выдачи результатов обследования. Оценки расходов должны быть максимально детализованными. Именно поэтому необходимо понимать в подробностях все этапы в ходе проведения обследования. В бюджете показаны затраты на персонал, оборудование и все прочие статьи расходов. Если лимит имеющихся в распоряжении средств определен заранее (а обычно так и бывает), общий бюджет обследования должен укладываться в эти рамки. Рекомендуется также при подготовке бюджета следовать общим рекомендациям финансирующей организации. Это может упростить одобрение проекта бюджета. Если возникает необходимость отклонения от предписанного бюджета, то требуется одобрение со стороны соответствующей(их) организации(й). Финансовые потребности для проведения обследования должны определяться на самой ранней стадии. Как правило, бюджет в значительной степени зависит от плана обследования, требуемой точности данных и географического охвата. На рисунке 2.2, ниже, показана одна из возможных ведомостей расходов.

15. Крайне важно создать эффективную систему контроля расходов в организации, проводящей обследование. В большинстве крупномасштабных обследований существует высокий риск утраты контроля над расходованием средств после начала работ на местах. В этих условиях значительные денежные суммы подчас направляются на цели, не связанные с основными операциями в рамках обследования. Продуманный контроль расходов помогает осуществлять мониторинг фактических затрат и сравнивать их со сметой расходов и фактически осуществленным объемом работ. Абсолютно необходимо, чтобы отвечающие за данное обследование руководители обеспечили отчетность по затраченным средствам. Это значительно повышает доверие к организации, проводящей обследование.

Рисунок 2.2
Примерная ведомость расходов на программу обследования домашних хозяйств

Мероприятия	Сметные единицы труда (человеко-месяцы, за исключением случаев, когда они обозначены иначе)	Издержки на единицу продукции (соответствующая единица валюты на человеко-месяц, за исключением случаев, когда они обозначены иначе)	Сметные общие затраты (соответствующая единица валюты)
I. **Планирование и подготовительные мероприятия**			
A. Начальное планирование и последующий контроль (руководящий персонал)			
B. Выбор и конкретизация темы			
1. Планирование темы			
2. Подготовка планов табулирования			
3. Делопроизводство и иные службы			
C. Разработка проекта обследования			
1. Начальное планирование проекта: структура обследования, охват населения, процедуры выборки, методы сбора данных и т. д. (персонал категории специалистов)			
2. Разработка материалов по выборке:			
a) Картографические материалы (предполагаются наличные материалы переписи): расходы на персонал карты и оборудование			
b) Списки домашних хозяйств на местах (2 000 счетных участков): расходы на персонал (главным образом регистраторы) путевые расходы			
c) Отбор и подготовка выборки на основе списков на местах			
D. Подготовка и печатание вопросников и других бланков			
1. Персонал категории специалистов			
2. Делопроизводство и иные службы			
3. Типографские расходы (после предварительных проверок)			
E. Предварительная проверка			
1. Планирование персонала категории специалистов:			
a) начальная подготовка			
b) анализ результатов и переработка материалов			
2. Контролеры на местах:			
a) расходы на персонал			
b) путевые расходы			

Мероприятия	Сметные единицы труда (человеко-месяцы, за исключением случаев, когда они обозначены иначе)	Издержки на единицу продукции (соответствующая единица валюты на человеко-месяц, за исключением случаев, когда они обозначены иначе)	Сметные общие затраты (соответствующая единица валюты)
3. Регистраторы:			
a) расходы на персонал			
b) путевые расходы			
F. Подготовка инструкций и учебных материалов для использования на местах			
1. Персонал категории специалистов			
2. Делопроизводство и иные службы			
3. Расходы на размножение документов			
G. Прочие работы по планированию (например, связь с общественностью и реклама)			
H. Предварительный итог по данной группе компонентов			
1. Руководящий персонал			
2. Персонал категории специалистов			
3. Технический персонал			
4. Состав категории обслуживания			
5. Путевые расходы			
6. Типографские расходы			
7. Картография и прочие расходы			
Предварительный итог			
II. Операции на местах			
A. Подготовка контролеров на местах			
1. Расходы на персонал			
2. Расходы на жилье и питание			
3. Путевые расходы			
B. Подготовка регистраторов			
1. Расходы на контролеров			
2. Расходы на регистраторов:			
a) расходы на персонал			
b) путевые расходы			
C. Сбор данных (включая контроль качества)			
1. Расходы на контролеров:			
2. Расходы на регистраторов			
a) расходы на персонал			
b) путевые расходы			
D. Администрация на местах			
1. Руководство на местах			
2. Путевые расходы			
3. Прочие расходы (например, на осуществление контроля и доставка материалов)			

Мероприятия	Сметные единицы труда (человеко-месяцы, за исключением случаев, когда они обозначены иначе)	Издержки на единицу продукции (соответствующая единица валюты на человеко-месяц, за исключением случаев, когда они обозначены иначе)	Сметные общие затраты (соответствующая единица валюты)
E. Предварительный итог по компонентам			
1. Персонал категории специалистов			
2. Технический персонал			
3. Персонал категории обслуживания			
4. Путевые расходы			
5. Командировочные расходы			
6. Проведение опросов			
7. Прочие расходы			
Предварительный итог			
III. Обработка данных			
A. Системное планирование			
B. Компьютерное программирование			
C. Техническое кодирование			
1. Начальное кодирование			
2. Контроль качества			
3. Проверка			
D. Операции кодирования на диск			
1. Начальный ввод данных			
2. Контроль качества			
3. Проверка			
E. Время использования компьютера (включая расходы на операторов и содержание компьютера)			
F. Прочие расходы на обработку данных (материалы и т.д.)			
G. Предварительный итог по компонентам			
1. Персонал категории специалистов			
2. Технический персонал			
3. Персонал, осуществляющий контроль качества			
4. Персонал категории обслуживания			
5. Расчеты			
6. Прочие расходы			
Предварительный итог			
IV. Подведение итогов и публикация данных			
A. Рабочее время специалистов			
B. Издательские расходы			
V. Руководство и координация (постоянный контроль всех мероприятий)			

	Мероприятия	Сметные единицы труда (человеко-месяцы, за исключением случаев, когда они обозначены иначе)	Издержки на единицу продукции (соответствующая единица валюты на человеко-месяц, за исключением случаев, когда они обозначены иначе)	Сметные общие затраты (соответствующая единица валюты)
VI.	Предварительный итог			
VII.	Оценочный анализ и методологические исследования (может исчисляться как 10 процентов нарастающего итога)			
VIII.	Общие накладные расходы (могут исчисляться как 15 процентов от нарастающего итога по административным расходам, аренде помещений, общему снабжению и т. п.)			
IX.	Итого			

Источник: Организация Объединенных Наций (1984 год).

2.2. Проведение обследований

2.2.1. Методы сбора данных

16. Существует целый ряд методов, используемых для сбора данных, к которым, в том числе, относятся: прямое наблюдение и измерение; рассылка вопросников по почте; проведение опросов по телефону и в личных беседах.

17. *Прямое наблюдение и измерение*. Прямое наблюдение и измерение представляют собой идеальный метод, так как эти способы сбора данных обычно более объективны. Не возникает проблем ни с провалами памяти, ни с необъективностью со стороны респондентов или регистраторов. Прямое наблюдение и измерение применяются, например, в следующих областях:

a) некоторые аспекты обследований потребления продуктов питания;

b) сбор информации о ценах, когда регистраторы могут закупать те или иные продукты и регистрировать их цены.

18. Этот метод, несмотря на свою полезность, отличается тем недостатком, что он требует больших затрат с точки зрения ресурсов и времени. В большинстве случаев регистраторы должны использовать то или иное оборудование. Как свидетельствует опыт, метод прямого наблюдения и измерения оказывается полезным и практичным при сравнительно небольших размерах выборки или совокупности.

19. *Рассылаемые по почте вопросники*. Метод рассылки вопросников по почте отличается сравнительной дешевизной и сжатыми сроками. Основной статьей расходов на этапе сбора данных являются почтовые сборы. После составления и печатания вопросников они рассылаются респондентам (т. е. людям, от которых ожидается заполнение данного вопросника). Считается, что респонденты грамотны, так как предполагается, что они самостоятельно будут заполнять вопросники. Такое допущение, однако, может оказаться ошибочным, особенно в развивающихся странах, где до сих пор отмечается низкий уровень грамотности. Основное уязвимое место этого

метода состоит в свойственной ему высокой доле неполучения ответов (т. е. большая доля лиц, не заполнивших вопросники и/или не ответивших на некоторые содержащиеся в них вопросы), что может быть обусловлено сложностью использованных вопросников. Нельзя также полностью исключить безразличие со стороны населения. В некоторых случаях присутствуют хорошие показатели возврата вопросников, но высокая доля неполучения ответов на отдельные вопросы.

20. В попытке повысить долю получения ответов на вопросы нереспондентам (т. е. лицам, не ответившим на вопросы обследования) могут рассылаться напоминания. Рекомендуется, однако, сформировать подвыборку нереспондентов и провести их опрос с помощью личной беседы. Необходимость в таких действиях может возникнуть, поскольку характеристики не ответивших домохозяйств могут кардинально отличаться от характеристик тех, которые ответили (см. обсуждение вопроса стратификации и его преимуществ в главе 5 и приложении I). В этом случае единицы, ответившие и не ответившие на вопросы, должны рассматриваться как две последующие стратификации в рамках административных единиц, которым при подготовке к получению оценок должны придаваться различные весовые коэффициенты (взвешивание результатов обследования более детально обсуждается в последующих главах, в частности в главе 6). В целях повышения доли ответивших на вопросы высылаемые по почте вопросники должны иметь привлекательный вид и быть короткими и по возможности простыми. Вложение в вопросники маркированных и снабженных адресом конвертов также может помочь повышению доли ответивших на вопросы.

21. Для успешного использования этого метода требуется также иметь по возможности оперативный обновленный инструментарий выборки. Следовательно, необходимо обновлять адреса респондентов. Проводящая обследование организация должна также удостовериться в том, что респонденты способны самостоятельно заполнить вопросники.

22. Ниже кратко изложены некоторые преимущества и недостатки обследований с помощью рассылки вопросников по почте:

Преимущества:

a) они дешевле;

b) выборка может быть разбросана по обширной территории;

c) исключены погрешности, обусловленные действиями регистраторов;

d) они проводятся в сжатые сроки.

Недостатки:

a) высокая доля неполучения ответов на вопросы;

b) ответы на вопросы принимаются сразу в их окончательном виде, поскольку нет возможности задать наводящие вопросы;

c) при обследовании мнений и отношений трудно удостовериться в том, отвечал ли респондент на вопросы без посторонней помощи;

d) этот метод полезно применять только тогда, когда вопросы носят сравнительно простой характер, и, следовательно, он неприемлем для комплексных обследований.

23. *Метод персональной беседы*. Это — наиболее распространенный в развивающихся странах метод сбора данных в рамках крупномасштабных выборочных обследований. Наряду с обычно высокой долей получения ответов в результате проведения персональных бесед, этот метод оптимален в силу высокого уровня неграмотности в некоторых из этих стран. Данный метод

предусматривает посещение регистраторами отдельных респондентов для сбора информации путем получения ответов на вопросы. Основным преимуществом этого метода является то, что регистраторы могут убедить респондентов (мотивируя их тем или иным образом) дать ответы на вопросы и разъяснить цели обследований. Кроме того, при использовании метода персональной беседы существует больше возможностей для сбора статистической информации по концептуально сложным вопросам, на которые при обследовании с помощью почтовой рассылки вопросников с большой вероятностью могут даваться двусмысленные ответы.

24. Тем не менее определенные недостатки присущи и методу персональной беседы. Например:

a) отдельные регистраторы могут по-разному толковать вопросы, искажая таким образом результаты обследования, поскольку лишь немногие регистраторы постоянно сверяются с инструкциями;

b) в процессе наводящей беседы некоторые регистраторы могут предлагать респондентам определенные ответы;

c) личные особенности регистраторов, например возраст, пол, а иногда и расовая принадлежность могут повлиять на отношение респондентов;

d) регистраторы могут неправильно зачитывать вопросы, поскольку их внимание разделено между ведением беседы и записью ответов.

25. В совокупности перечисленные выше недостатки являются основным источником так называемых «погрешностей, обусловленных действиями регистраторов», поскольку, как показали исследования, могут приводить в обследованиях к серьезным ошибкам регистрации получаемых данных (проблемы ошибок регистрации данных обследования рассматриваются в главе 8).

26. При опросе респондентов следует принимать во внимание следующие моменты:

a) регистратор должен ясно понимать цель каждого вопроса, как она разъясняется в пособии для регистраторов. Важно, чтобы регистраторы постоянно сверялись с пособием;

b) как показывает опыт, регистратору лучше следовать последовательности вопросов в вопроснике. В большинстве вопросников значительное внимание уделяется порядку расположения вопросов с учетом мотивации респондентов, взаимоувязки рассматриваемых тем, облегчения воспоминаний респондента о прошлых событиях и наиболее деликатных вопросов;

c) регистраторы должны любыми способами воздерживаться от предложения ответов респондентам;

d) должны быть заданы все вопросы. Таким образом, сводится к минимуму неполучение ответов на отдельные вопросы. Кроме того, ни один из пунктов вопросника не должен оставаться незаполненным, если только это не соответствует порядку пропуска тех или иных пунктов. Если тот или иной вопрос не относится к определенному респонденту, то должно быть сделано соответствующее примечание. Такой подход служит гарантией для руководителя обследования, что обработаны все вопросы, включенные в вопросник.

2.2.2. Разработка вопросника

27. Как только определены цели обследования и план табулирования, может быть разработан соответствующий вопросник. Вопросник играет центральную роль в процессе обследования, в рамках которого информация передается от ее носителей (респондентов) к тем, кому она необхо-

дима (пользователям). Это — именно тот инструмент, посредством которого информационные потребности пользователей выражаются в оперативном виде, а также основной источник данных, вводимых в систему их обработки для определенного обследования.

28. Размер и формат вопросника требуют весьма серьезного рассмотрения. Рекомендуется разрабатывать структуру вопросника в период планирования обследования. Если вопросники нужно рассылать респондентам, они должны отличаться привлекательностью и простотой. Это может увеличить долю полученных ответов. С другой стороны, тот вопросник, который будет использоваться для записи ответов регистраторами на местах, должен быть достаточно прочным, чтобы выдержать активную работу с ним.

29. В идеальном случае вопросник должен быть составлен таким образом, чтобы облегчить сбор актуальных и достоверных данных. Для повышения точности данных обследования, особое внимание необходимо уделить расположению и последовательности пунктов в вопроснике и их формулировке. Респондент должен быть заинтересован в заполнении этого документа. Вопросник должен иметь достаточно большой формат для легкости чтения вопросов как респондентом, так и регистратором. Ни в коем случае нельзя упускать из виду то, что каждый вопросник должен содержать четкие инструкции.

30. Таким образом, проводящей обследование группе необходимо уделять особое внимание составлению точных определений подлежащих сбору данных, а также точных спецификаций касательно перевода информационных требований и связанных с ними концепций в операционные процедуры. В этой связи обычным и, как правило, необходимым мероприятием становится предварительное тестирование вопросника, если он уже не был в полной мере опробован в предшествующих обследованиях.

31. Итак, хороший вопросник должен:

a) обеспечивать возможность сбора точной информации для своевременного удовлетворения потребностей потенциальных пользователей данных;

b) облегчать работу по сбору и обработке данных и составлению таблиц;

c) обеспечивать «экономичность» процесса сбора данных, а именно избегать сбора любой несущественной информации;

d) позволять проведение всеобъемлющего и содержательного анализа и целенаправленного использования собранных данных.

32. Это означает разработку такого вопросника обследования, который давал бы информацию максимально достижимого качества с особым упором на ее актуальность, своевременность и точность. Для эффективного осуществления этого процесса необходимо свести к минимуму расходы и рабочую нагрузку при сборе необходимой информации.

2.2.2.1. Структура вопросов

33. В вопросниках для выборочных обследований используются «открытые» и «закрытые» вопросы. На открытый вопрос респондент дает свой собственный ответ. В обследовании мнений и отношений респондентам может быть предложено сформулировать, что они считают хорошим качеством жизни. Очевидно, что разные респонденты дадут свое собственное определение того, что они вкладывают в понятие хорошего качества жизни. С другой стороны, закрытый вопрос ограничивает респондента в выборе ответов списком, уже определенным группой, проводящей обследование. Ниже приводятся примеры ответов на закрытые вопросы:

Имеете ли вы какие-либо хронические психические расстройства, ограничивающие вашу повседневную деятельность?

❑ Да ❑ Нет

Как вы оцениваете состояние вашего зрения (даже в очках или контактных линзах, если вы их используете)?

1. ❑ слепота
2. ❑ серьезные хронические проблемы
3. ❑ некоторые хронические проблемы
4. ❑ нет никаких проблем

34. Преимущества использования закрытых вопросов заключаются в том, что они: *a*) дают более единообразные ответы и *b*) просты в обработке. Основным недостатком таких вопросов является то, что возможные ответы должны быть сформулированы разработчиком обследования. В этом случае могут быть упущены важные возможные ответы. В большинстве обследований сложные темы и проблемы, связанные с мнениями и отношениями, которые могут быть неизвестны, наилучшим образом рассматриваются с помощью открытых вопросов.

2.2.2.2. Формулирование вопросов

35. Вопросы должны быть ясными, точными и недвусмысленными. Респондент не должен гадать, какую информацию хочет от него/нее получить регистратор. Используемые определения и понятия могут казаться очевидными для руководителя обследования, но не для респондента. Вследствие этого респондент может при ответах на вопросы руководствоваться собственными соображениями. Конечным результатом будет распространение ошибок регистрации данных обследования. Рассмотрим простой пример. Вопрос: «каков ваш домашний адрес» создает путаницу во многих африканских странах, особенно среди городского населения, если не дать четкое определение понятия «дом». Некоторые респонденты полагают, что «дом» означает деревню, откуда они родом.

2.2.2.3. Провокационные вопросы

36. Так называемый провокационный вопрос побуждает респондента ответить на вопрос определенным образом. Это означает, что вопрос может быть в некоторой степени смещен в пользу определенного ответа. Вот пример провокационного вопроса в обследовании сферы здравоохранения: «Сколько раз в неделю вы выпиваете более двух бутылок пива?». Этот вопрос вынуждает респондента признать, что он/она пьет пиво и фактически не менее двух бутылок в день. Такие вопросы имеют тенденцию искажать ответы респондентов. Важно избегать «формирования» данных: целью является лишь их сбор.

2.2.2.4. Актуальность вопросов

37. Задачей вопросника является извлечение информации, которая будет использована для изучения ситуации. Таким образом, проводящей обследование организации абсолютно необходимо задавать актуальные вопросы для получения истинной картины о положении дел в конкретной исследуемой области. Вопросы, включенные в вопросник, должны быть актуальными для большинства респондентов. Например, бессмысленно в сельской местности, типичной сегодня для

большинства африканских стран, использовать вопросник, загроможденный вопросами о личных достижениях по части высшего (университетского) образования. Аналогичным образом, не следует включать в обследование уровня рождаемости лиц женского пола в возрасте, например, 10 лет и моложе и задавать им вопросы о числе когда-либо рожденных ими детей, о том замужем ли они, разведены или являются вдовами. Такие вопросы будут актуальны для женщин старше определенного возраста, а не для девочек моложе детородного возраста.

2.2.2.5. Последовательность вопросов

38. Пункты вопросника должны располагаться в такой последовательности, чтобы заинтересовать респондента и облегчить его/ее обращение к памяти, а также оказать ему/ей помощь в выдаче точной информации. Рекомендуется, чтобы первые вопросы были простыми, интересными и не деликатными. Это укрепляет доверие респондента, позволяя ему/ей пройти весь опрос целиком, на который он/она в большинстве случаев соглашается добровольно. Сравнительно стандартной практикой обследований домашних хозяйств стало начинать общую последовательность вопросов с тех из них, которые идентифицируют единицу выборки, таких как адрес, за которыми следуют вопросы касательно описания домашнего хозяйства и проживающих в нем лиц, включая, например, демографические характеристики. И наконец, за этим задаются вопросы, составляющие основной предмет обследования[1]. Как правило, деликатные вопросы включаются в число вопросов, которые задаются в последнюю очередь. Следует подчеркнуть, что между вопросами должна присутствовать логическая связь, особенно между теми, которые взаимозависимы.

2.2.3. План табулирования и анализа

39. Существует удобный метод, помогающий разработчику обследования добиться более высокой точности от тех инструментов, которые предназначены для удовлетворения потребностей пользователей в информации, отраженной в списке вопросов или в целях обследования. Этот метод заключается в разработке планов табулирования и создании макетов таблиц. Макеты таблицы — это проекты таблиц, включающие в себя все, кроме фактических данных. Схема табулирования должна, как минимум, указывать названия строк и столбцов таблицы, обозначая основные переменные величины, которые будут табулированы, справочные переменные, которые должны использоваться для классификации, и группы совокупности (цели обследования, элементы или единицы), к которым относятся различные таблицы. Желательно также по возможности максимально подробно показать категории классификации, хотя их можно скорректировать и позднее, когда будет более точно известно распределение выборки по категориям ответов.

40. Важность плана табулирования можно рассматривать с нескольких точек зрения. Например, созданные макеты таблицы укажут на то, позволят ли собранные данные составить пригодные для использования таблицы. Они укажут не только на пробелы, но и на лишние позиции. Более того, дополнительное время, затраченное на составление макетов таблиц, с лихвой окупается на этапах табулирования данных, сокращая затраты времени на разработку и составление фактических таблиц.

41. Должна также учитываться тесная взаимосвязь между планом табулирования и принятым для данного обследования планом выборки. Например, географическая разбивка таблиц возможна только в том случае, если такую разбивку предусматривает разработанный план выборки. Кроме того, размер выборки может вызвать необходимость в ограничении числа ячеек в перекрест-

1 См. Организация Объединенных Наций, 1984 год.

ных таблицах во избежание составления чересчур разреженных таблиц. Возможно, что иногда такой план должен подвергаться модификации в ходе самой работы по табулированию данных. Может возникнуть необходимость в объединении категорий для сокращения числа пустых ячеек; или интересные выводы, обнаруженные в первичных данных, могут обусловить необходимость составления новых таблиц. Тот способ, посредством которого которым собранные в ходе обследования домохозяйств данные будут использованы для ответов на вопросы (для достижения целей) обследования, в общем можно назвать «планом анализа данных». Такой план содержит детальное разъяснение тех данных, которые необходимы для достижения целей обследования. Разработчики плана обследования должны постоянно сверяться с этим планом при работе над детализацией вопросника обследования. Пожалуй, не вызывает сомнения то, что план анализа должен являться основным ориентиром в плане выдачи рекомендаций по анализу результатов обследования.

2.2.4. Проведение работ на местах

42. В большинстве развивающихся стран проведение работ на местах зачастую существенно ограничено недостатком ресурсов. Тем не менее, если проведение обследования необходимо, то работы на местах должны быть надлежащим образом организованы и выполнены таким образом, чтобы эффективно использовать имеющиеся в распоряжении группы обследования ограниченные ресурсы. Для успеха работ по проведению обследования те лица, которые заняты планированием этих работ, должны ясно понимать концептуальные аспекты предмета обследования. Кроме того, регистраторы должны в совершенстве овладеть практическими процедурами, что обеспечит успешный сбор точных данных. Для успешного осуществления работ по проведению обследования всегда необходимо обеспечивать хорошую и эффективную организацию работ на местах.

2.2.4.1. Оборудование и материалы

43. Во многих развивающихся странах необходимо задолго до обследования обеспечить наличие и привести в рабочее состояние такое оборудование, как автомобили, катера, велосипеды и т. д. Необходимо также иметь определенное количество запчастей. Автомобили и велосипеды обеспечивают мобильность руководства группы, а также контролеров/регистраторов.

44. В нужном количестве должны иметься такие расходные материалы, как папки, скоросшиватели, карандаши, точилки, блокноты и топливо (для автомобилей) для использования в работах по проведению обследования.

2.2.4.2. Управление работами по проведению обследования

45. Крупномасштабное выборочное обследование обычно представляет собой сложное мероприятие, требующее больших затрат сил. Вследствие этого нельзя недооценивать потребность в осмотрительном, эффективном и действенном руководстве работами на различных уровнях.

46. Должна присутствовать ясная и четко определенная цепочка инстанций от руководителя обследования до регистратора. Следует отметить, что была доказана полезность использования контрольных формуляров для мониторинга хода обследования.

2.2.4.3. Рекламно-пропагандистская деятельность

47. Некоторые обследования лишь частично достигают своих целей из-за высокой доли неполучения ответов в связи с отказом от участия в обследовании. Вследствие этого организаторам обследования неминуемо придется провести в целях его успеха ряд информационно-пропагандистских

кампаний. Как показывает опыт, пропагандистская деятельность играет важную роль в привлечении респондентов к сотрудничеству, даже несмотря на то, что некоторые организации/агентства рассматривают расходы на такую деятельность как напрасную трату ресурсов.

48. К пропагандистской деятельности применимы различные подходы, зависящие от сложившихся обстоятельств. Например, в городских районах некоторых стран рекламно-информационные плакаты могут дополняться сообщениями по радио, телевидению и в газетах, в то время как в сельской местности могут использоваться только плакаты и радиосообщения.

49. Наряду с этим может возникнуть необходимость в организации встреч с местными авторитетными лидерами в отдельных районах. В ходе таких встреч люди должны быть проинформированы о целях обследования. Кроме того, надлежит обратиться с просьбой к местным лидерам, чтобы они убедили жителей своего района предоставлять регистраторам требуемую информацию.

50. До начала работ на местах важно опубликовать соответствующие правовые основания для проведения обследования. В объявлении наряду с прочей информацией должны быть указаны цели и продолжительность обследования, а также охватываемые им темы.

2.2.4.4. Отбор регистраторов

51. Регистратор служит связующим звеном с респондентами. Тот факт, что он/она является представителем проводящей обследование организации, находящимся в постоянном контакте с респондентом, явственно указывает на то, почему работа регистратора столь важна для успеха программы обследования. В связи с этим отбору регистраторов следует уделять особое внимание, и такой отбор необходимо осуществлять весьма тщательно. Регистратор должен быть способен эффективно общаться с респондентом и обладать необходимыми качествами для сбора точной информации в разумные сроки.

52. В зависимости от вида обследования регистратор должен иметь соответствующий уровень образования. В дополнение к этому регистратор должен быть способен честно фиксировать получаемую информацию, не «фальсифицируя цифры». Отобранные регистраторы обязаны следовать инструкциям и использовать определения и понятия, указанные в предназначенном для них учебном пособии по работам на местах.

53. Следующие процедуры могут пригодиться при отборе регистраторов:

a) соискатели на должности регистраторов должны заполнить заявление, указав в нем свой возраст, семейное положение, фактический адрес, уровень образования и места прежней работы;

b) прошедшие первичный отбор кандидаты должны пройти тест оценки интеллекта и дополнительный тест на простые арифметические расчеты;

c) поскольку наряду с письменными тестами обычно возникает необходимость в собеседовании с соискателями, такие собеседования должны проводиться комиссией, которая дает кандидатам независимую оценку; некоторыми из качеств, оцениваемых у кандидатов, являются дружелюбие, интерес к работе, самовыражение и внимательность.

54. Работы на местах могут быть утомительным занятием, включая трудности передвижения по пересеченной местности; и в связи с этим отобранные регистраторы должны быть преданы своему делу и готовы работать в сложных условиях.

2.2.4.5. Подготовка регистраторов

55. До направления на работу на местах отобранные регистраторы должны пройти всестороннюю подготовку. Основной целью программы подготовки является обеспечение единообразия в процедурах проведения опроса в рамках обследования. Это, безусловно, необходимо для предотвращения различий в толковании регистраторами определений, понятий и целей обследования и, следовательно, для сведения к минимуму погрешностей, обусловленных действиями регистраторов.

56. Ответственными за подготовку должны быть квалифицированные инструкторы. Такие инструкторы, безусловно, должны быть прекрасно осведомлены о целях и задачах обследования. Предпочтительно, чтобы они входили в состав группы, проводящей данное обследование.

57. Регистраторы должны быть тщательно проинструктированы относительно целей обследования и относительно того, каким образом будут использоваться его результаты. Для надлежащей оценки регистраторами целей обследования они должны быть хорошо подготовлены в вопросах понятий и определений, используемых в вопроснике.

58. В рамках учебного процесса регистраторы в присутствии инструктора должны по очереди разъяснять своим сокурсникам различные пункты вопросника. Должны быть организованы практические семинары как в классе, так и в фактических условиях на местах. Например, регистраторы-стажеры могли бы поочередно задавать друг другу вопросы в ходе классных занятий, а затем совершать выезды в окрестные районы, где проводить тренировочные опросы нескольких домашних хозяйств. Всегда должен присутствовать инструктор для руководства действиями регистраторов и исправления их ошибок. После проведения опросов на местах регистраторы-стажеры должны обсуждать полученные результаты под руководством инструктора. По итогам программы подготовки руководитель обследования должен принять решение о том, требуется ли дополнительная подготовка для каких-либо регистраторов, и есть ли среди них лица, полностью непригодные для такой работы.

2.2.4.6. Контроль за работами на местах

59. Общепризнано, что подготовка является необходимым условием эффективной и успешной работы на местах. Тем не менее подготовка без надлежащего контроля может не принести желаемых результатов. Успешное проведение работ на местах требует непрерывного, целенаправленного и эффективного контроля со стороны руководящих сотрудников, имеющих больший опыт и более высокую квалификацию, чем регистраторы. Контролеры должны пройти подготовку по всем аспектам проведения обследования. Нельзя переоценивать тот факт, что контролер является важным связующим звеном между организацией, ведущей сбор данных, и регистратором. От контролера требуется организовать работу регистраторов, выдавая им конкретные задания и определяя районы работы; он или она изучает проделанную работу и поддерживает высокий уровень приверженности выполнению программы обследования среди регистраторов. Мы рекомендуем поддерживать по возможности сравнительно высокое соотношение между контролирующим персоналом и регистраторами. В качестве эталона для большинства обследований домашних хозяйств предложено соотношение — один контролер на четыре-пять регистраторов. Тем не менее это — всего лишь рекомендация.

2.2.4.7. Последующая работа с лицами, не ответившими на вопросы

60. В большинстве обследований могут иметь место случаи неполучения ответов (см. главу 8 касательно ошибок регистрации данных обследования). Некоторые респонденты могут отказать-

ся сотрудничать с регистраторами; в некоторых случаях могут быть пропущены определенные пункты вопросника. В случае информирования контролера о наличии той или иной единицы, не отвечающей на вопросы, он/она должен связаться с этой единицей выборки и постараться получить информацию, пользуясь своей более высокой квалификацией и бóльшим опытом. Поскольку оперативной целью любого обследования является достижение максимально возможной доли ответивших на вопросы, рекомендуется собрать информацию от подвыборки первоначально не ответивших респондентов. В этом случае усилия при проведении обследования перенаправляются на работу с этой подвыборкой с предпочтительным использованием контролеров в качестве регистраторов.

2.2.4.8. Сокращение показателей неполучения ответов

61. При разработке и проведении обследования домашних хозяйств важно разработать надлежащие процедуры обследования, направленные на максимальное увеличение доли ответивших на вопросы. Мы хотели бы подчеркнуть важность наличия процедур по снижению числа отказов, таких как организация повторного посещения для проведения опроса в удобное для респондента время. Кроме того, цели и области использования результатов обследования должны тщательно разъясняться колеблющимся респондентам для привлечения их к сотрудничеству. Гарантии конфиденциальности также могут содействовать снижению опасений респондентов по поводу своих ответов, которые могут быть использованы в иных целях, нежели заявленные в обследовании.

62. Если дома не удалось застать никого из жильцов, то должны быть организованы повторные посещения в разное время дня. Рекомендуется проводить не менее четырех повторных посещений.

63. Важно избегать проблем, связанных с невозможностью найти отобранные единицы выборки, что может стать серьезным источником неполучения ответов. Эта проблема решается с помощью использования максимально обновленного инструментария выборки (см. главу 4, в которой подробно рассматривается эта тема).

Справочная литература и дополнительные источники информации

Kiregyera, B. (1999). *Sample Surveys: With Special Reference to Africa.* Kampala: PHIDAM Enterprises.

Statistics Canada (2003). *Survey Methods and Practices.* Ottawa.

United Nations (1982). Preliminary version. National Household Survey Capability Programme: DP/UN/INT-81-041/2 non-sampling errors in household surveys: sources, assessment and control. New York:

Организация Объединенных Наций (1984). *Руководство по обследованию домашних хозяйств* (Переработанное издание). Методологические исследования, № 31; в продаже под № R.83. XVIII.13.

____(1998). Принципы и рекомендации по проведению переписей населения и жилого фонда. (Первый пересмотренный вариант). Статистические документы, № 67/Rev.1. В продаже под № R.98. XVII.8.

United Nations (2002). Technical report on collection of economic characteristics in population censuses. New York and Geneva: ST/ESA/STAT/119. English only. Statistics Division, Department of Social and Economic Affairs, and Bureau of Statistics, International Labour Office.

Zanutto, E., and A. Zaslavsky (2002) Using administrative records to improve small area estimation: an example from the U.S. Decennial Census. *Journal of Official Statistic* (Statistics Sweden), vol. 18, No. 4, pp. 559–576.

Глава 3

Стратегии формирования выборки

3.1. Введение

1. В то время как в главе 2 по теме планирования обследований дан общий обзор различных этапов процесса обследования домашних хозяйств, данная глава — это первая из нескольких глав, в которых главный акцент сделан исключительно на тех или иных аспектах выборки, т. е. на основном предмете настоящего руководства. В этой главе дается краткое сравнение вероятностной и детерминированной выборки, а также приводятся аргументы в пользу того, почему последняя должна всегда использоваться в обследованиях домашних хозяйств. Большое внимание уделяется размеру выборки: многим определяющим ее параметрам и способам их расчета. Представлены методы достижения эффективности выборки в обследованиях домашних хозяйств. К ним относятся стратификация, гнездовая выборка и поэтапная выборка с особым упором на двухэтапные планы выборки (см. определения и описания этих понятий в таблице 3.1 и приложении I). Предоставлены различные варианты формирования выборки, а также дается подробное описание двух основных планов выборки, которые были использованы во многих странах. Кроме того, рассмотрены такие специальные темы, как *a*) формирование выборки в два этапа для охвата «редких» групп населения и *b*) создание выборки для выявления изменений или трендов. Глава завершается сводным изложением рекомендаций.

3.1.1. Обзор

2. Практически все планы выборки для обследований домашних хозяйств, проводимых как в развитых, так и в развивающихся странах, являются комплексными в силу таких характеристик, как многоэтапность, стратификация и гнездовая выборка. Кроме того, их сложность возрастает в связи с тем фактом, что обследования домашних хозяйств на национальном уровне часто имеют общий характер, охватывая многочисленные темы, интересующие правительство. Поэтому в настоящем руководстве основное внимание уделяется многоэтапным стратегиям выборки.

3. Для получения желаемого результата разработанный надлежащим образом план выборки для обследования домашних хозяйств должен, как симфония, гармонично объединять многочисленные элементы. Формирование выборки должно осуществляться *поэтапно* с тем, чтобы эффективно выявить места проведения опросов и выбрать домашние хозяйства. План выборки должен быть *стратифицирован* таким образом, чтобы гарантировать, что фактически сформированная выборка равномерно распределяется по небольшим географическим районам и подгруппам населения. В плане выборки должны использоваться *кластеры*, обычно представляющие собой географически определенные единицы, из которых отбираются домашние хозяйства, в целях сохранения затрат на управляемом уровне. В то же время следует избегать излишней разбивки на *кластеры*, поскольку план с чересчур мелкой *гнездовой выборкой* отрицательно влияет на надежность данных (формирование кластеров рассматривается в разделе 3.3.5). *Размер*

выборки должен учитывать противоречащие друг другу потребности, с тем чтобы были оптимально сбалансированы уровни затрат и точность собираемых данных. Размер выборки должен также соответствовать насущным потребностям пользователей, желающих получить данные по областям обследования, а именно — по подгруппам населения или небольшим районам. План выборки должен предусматривать достижение максимальной точности двумя основными способами: во-первых, используемый (или формируемый) *инструментарий выборки* должен быть как можно более полным, точным и обновленным и, во-вторых, должны использоваться такие методы формирования выборки, которые сводят к минимуму непреднамеренные погрешности, которые иногда допускаются ее составителями. План выборки также должен поддаваться самооценке, иными словами план выборки должен быть таким, чтобы *ошибки выборки* могли быть оценены и давали возможность пользователям рассчитывать надежность основных результатов обследования. Ошибки выборки возникают в результате оценки характеристик той или иной совокупности, базирующихся на данных только о ее части, а не обо всей совокупности в целом.

4. Основной целью любого обследования является возможность делать заключения касательно исследуемой совокупности, базируясь на случайной выборке. Для достижения этой цели исследователь обычно стремится оценить некоторые неизвестные характеристики совокупности. В число общих подлежащих оценке характеристик/параметров совокупности входят суммарные величины, средние величины, соотношения и вариантности. Например, если $Y_1, Y_2, Y_3, \ldots \ldots Y_N$ являются значениями переменной *y* в совокупности, тогда

среднее значение для совокупности составляет $\bar{Y} = \frac{1}{N}\sum Y_i;$ (3.1)

вариантность совокупности составляет $\sigma^2 = \frac{1}{N}\left(\sum Y^2 - N\bar{Y}^2\right).$ (3.2)

В большинстве случаев оценки выборки используются и для оценки параметров совокупности. Например, среднее значение и вариантность выборки размера *n* для простой случайной выборки с замещением рассчитываются по формулам:

$$\bar{y} = \frac{1}{n}\sum y_i; \quad (3.3)$$

$$s^2 = \frac{1}{n-1}\left(\sum y_i^2 - n\bar{y}^2\right), \quad (3.4)$$

где $y_1, y_2, y_3, \ldots \ldots y_n$ — это значения переменной *y* для *n* числа единиц в выборке. В выборочных обследованиях исследователи рассчитывают вариантность отдельных случайных переменных для определения значения ошибки выборки в полученной оценки (см. определение ошибки выборки в таблице 3.1, более подробно ошибки выборки рассматриваются в главе 7 и приложении I). К факторам, влияющим на размеры вариантности выборки, относятся: однородность исследуемых переменных, размер выборки и план выборки (эти аспекты рассмотрены в различных разделах данной главы, главы 7, а основные принципы формирования выборки обследования представлены в приложении I).

5. В главах 3 и 4 подробно рассматривается каждый из параметров, используемых при формировании надлежащей выборки для обследований домашних хозяйств. Как правило, речь в основном идет о национальных обследованиях, хотя все описанные здесь методы применимы

и к крупным обследованиям на субнациональном уровне, например, ограниченным одним или несколькими регионами, провинциями, районами или городами. В силу огромной важности инструментария выборки для достижения приемлемых практических результатов в формировании выборки, глава 4 полностью посвящена этому вопросу.

3.1.2. Глоссарий терминов по выборке и связанным с ней областям

6. Мы начинаем с глоссария терминов, используемых в данной и последующих главах (см. таблицу 3.1). Этот глоссарий не предназначен для предоставления официальных определений терминов по выборке, частично относящихся к области математики. Вместо этого в нем дается описание терминов в контексте данного руководства, конечно, с упором на их применение в обследованиях домашних хозяйств.

Таблица 3.1
Глоссарий терминов по выборке и связанным с ней областям

Термин	Описание
Алгоритм оценки	Для данного плана выборки алгоритм оценки — это метод оценки параметра совокупности по данным выборки, например алгоритмом оценки является среднее арифметическое значение выборки
Быстрый подсчет	Означает операцию обновления данных, когда жилые помещения подсчиатываются приблизительно для получения текущего показателя размера: см. также *сбор статистических данных*
Вариантность выборки	Квадрат среднеквадратической ошибки или ошибки выборки
Вероятностная выборка	Методология отбора, согласно которой каждая единица совокупности (человек, домашнее хозяйство и т. д.) имеет известную, не равную нулю вероятность включения в выборку
Вес	Величина, обратная вероятности выбора; фактор повышения, применяемый к необработанным данным; также известен как *вес схемы*
Внутриклассовая корреляция	С помощью коэффициента внутриклассовой корреляции измеряется однородность схожих элементов
Выборка для определения трендов	План выборки для оценки изменений между различными периодами времени
Выборка по методу равновероятного выбора (*Epsem*)	Равновероятностная выборка
Выборка по стадиям, также известная как двойная выборка или выборка с последующей стратификацией	Формирование выборки (как правило) в две стадии по времени, при этом выборка на второй стадии обычно является подвыборкой выборки первой стадии; нельзя путать *с выборкой для определения трендов* (см. выше)
Выборка с вероятностью, пропорциональной размеру (ВПР)	Отбор единиц первого, (второго и т. д.) этапа, при котором каждая единица отбирается с вероятностью, пропорциональной ее показателю размера; см. также в тексте выборку с вероятностью, пропорциональной *предлагаемому* размеру (вппр)
Гнездовая выборка	Формирование выборки, в котором предпоследний этап предусматривает использование определенной территориальной единицы, такой как счетный район переписи
Детерминированная выборка	См. в разделе 3.2.2 текстовые описания примеров данного метода: выборка по квотам, по экспертной оценке, преднамеренная выборка, нерепрезентативная выборка, выборка путем случайного «блуждания»

Термин	Описание
Доверительная вероятность	Означает уровень статистической достоверности, с помощью которого определяется точность или допустимый предел погрешности оценок обследования. Стандартным значением, как правило, считается 95-процентный доверительный интервал
Доля выборки	Отношение размера выборки к общему количеству единиц совокупности
Инструментарий(и) выборки	Сборник материалов, из которых фактически формируется выборка, такой как список или набор районов, иными словами, набор единиц совокупности
Компактный кластер	Кластер выборки, состоящий из территориально прилегающих друг к другу домашних хозяйств
Комплексный план выборки	Означает использование нескольких этапов, формирование кластеров и стратификацию в выборке обследования домашних хозяйств в противоположность простой случайной выборке
Надежность (точность, допустимый предел ошибки)	Обозначает уровень ошибки выборки, связанной с данной оценкой обследования
Некомпактный кластер	Кластер выборки, состоящий из территориально рассредоточенных домашних хозяйств
Область обследования	Географическая единица, для которой предоставляются отдельные оценки обследования
Обследуемое население	Определение населения, которое предполагается охватить обследованием, также известно как *охватываемая генеральная совокупность*
Относительная среднеквадратическая ошибка (вариационный коэффициент)	Среднеквадратическая ошибка как процентная доля оценки обследования, иными словами, среднеквадратическая ошибка, деленная на оценку
Ошибка выборки (среднеквадратическая ошибка)	Случайная ошибка в оценке обследования в связи с тем, что проводится обследование выборки, а не всей совокупности; квадратный корень величины вариантности выборки
Ошибка регистрации	Погрешности оценок обследования, возникающие в связи с ошибками в схеме и проведении обследования; относится к *правильности* и *достоверности* оценки в отличие от ее надежности или точности
Первичная единица выборки (ПЕВ)	Территориально определенная административная единица, отобранная на первом этапе выборки
Показатель размера	При многоступенчатой выборке расчет или оценка размера (например, число лиц) в каждой единице на данном этапе
Поэтапная выборка	Методика, с помощью которой выборка административных районов и домохозяйств/лиц формируется последовательными этапами для выделения географических мест проведения обследования
Правильность (достоверность)	См. ошибка регистрации данных обследования
ПСВ	Простая случайная выборка (редко используется в обследованиях домашних хозяйств)
Размер выборки	Число отобранных домашних хозяйств или лиц
Размер кластера	(Среднее) количество единиц выборки — лиц или домашних хозяйств — в кластере
Самовзвешивание	План выборки, в котором все позиции имеют одинаковый вес в обследовании
Сбор статистических данных	Метод «охвата» того или иного географического района для определения адресов жилых помещений и/или домашних хозяйств, обычно применяемый в рамках мероприятий по обновлению инструментария выборки

Термин	Описание
Сегмент	Подразделение более крупного кластера, нанесенное на карту с указанием границ
Систематическая выборка	Отбор из списка, используя случайную точку отсчета и заранее определенный и последовательно применяемый интервал отбора
Скрытая стратификация	Означает стратификацию посредством территориального упорядочения инструментария выборки вместе с систематическим формированием выборки с вероятностью, пропорциональной размеру
Списочное формирование выборки	Осуществление отбора из списка единиц, входящих в инструментарий выборки
Стратифицированная выборка	Методология сведения инструментария выборки в подгруппы, являющиеся внутренне однородными при внешней разнородности, для обеспечения того, чтобы сформированная выборка была надлежащим образом «рассредоточена» по важнейшим подгруппам населения
Субсегментация (разбивка на фрагменты)	Как правило, проводимое на местах мероприятие, в рамках которого непредвиденно сформированные крупные кластеры разделяются на части для снижения объема работ по составлению списков
Территориальная выборка	Отбор географических единиц, входящих в инструментарий выборки (может включать в себя отбор *сегментов* территорий, определяемых как нанесенные на карту подразделения административных единиц)
Фиктивный этап отбора	Искусственно введенный этап отбора, предназначенный для упрощения осуществляемой вручную работы по выявлению подрайонов, в которых в конечном счете будут находиться кластеры выборки
Формирование кластеров; сгруппированный кластер	Относится к тенденции единиц выборки — лиц или домашних хозяйств — обладать аналогичными характеристиками
Эталонная выборка	«Супер» выборка, предназначенная для использования в нескольких обследованиях и/или в нескольких раундах одного и того же обследования, обычно в течение 10-летнего периода
Эффект схемы (*deff*)	Соотношение вариантности комплексного плана выборки к вариантности плана простой случайной выборки при выборке того же размера; иногда упоминается как *формирование кластеров*, хотя *deff*, помимо формирования кластеров, включает в себя эффект стратификации

3.1.3. Символы

7. В этой и следующих главах настоящего руководства используются стандартные символы (см. таблицу 3.2). Как правило, прописные буквы обозначают величины совокупности, а строчные буквы обозначают наблюдения выборки. Например, символы $Y_1, Y_2, Y_3, \ldots \ldots Y_N$ обозначают величины совокупности, в то время как символы $y_1, y_2, y_3, \ldots \ldots y_n$ обычно используются для указания выборочных значений. Из вышесказанного вытекает, что N — это размер совокупности, в то время как n обозначает размер выборки. Важно отметить, что параметры совокупности обозначаются либо прописными буквами английского алфавита, либо греческими буквами. Например, $\bar{Y}$ и σ обозначают среднее значение и среднеквадратическое отклонение совокупности, соответственно. Оценки параметров совокупности имеют знак ^, расположенный над прописной буквой, например $\overline{Y}$, в то время как символ выборочных значений $\bar{y}$ обозначается строчной буквой.

Таблица 3.2

Отдельные символы, используемые для обозначения величин совокупности и параметров выборки

Характеристики	Символы		
		Выборка	
	Совокупность	Оценки со знаком «^» над символом	Использование строчной буквы
Единицы	N	n	n
Наблюдения	$Y_1, Y_2, \ldots \ldots Y_i, \ldots \ldots Y_N$	$\hat{Y}_1, \hat{Y}_2, \ldots \ldots \hat{Y}_i, \ldots \ldots \hat{Y}_n$	$y_1, y_2, \ldots \ldots y_i, \ldots \ldots y_n$
Среднее значение	$\bar{Y}$	$\hat{\bar{Y}}$	$\bar{y}$
Пропорция	P	$\hat{P}$	p
Параметр и оценка	Θ	$\hat{\Theta}$	$\hat{\theta}$
Вариантность y	$\sigma^2(y)$	$\hat{\sigma}^2(y)$	$s^2(y)$
Среднеквадратическое отклонение y	$\sigma(y)$	$\hat{\sigma}(y)$	$s(y)$
Внутриклассовая корреляция	δ	$\hat{\delta}$	$\hat{\delta}$
Отношение	R	$\hat{R}$	r
Суммирование	$\sum_{i=1}^{N}$	$\sum_{i=1}^{n}$	$\sum_{i=1}^{n}$

3.2. Сравнение вероятностной выборки с другими методами выборки для обследований домашних хозяйств

8. Хотя обсуждение теории вероятностей лежит за рамками данного руководства, важно разъяснить, почему вероятностные методы играют незаменимую роль в формировании выборки для обследований домашних хозяйств. В данном разделе даются краткое определение и описание вероятностной выборки, а также излагаются причины, указывающие на важность этого параметра. Кратко упоминаются также другие методы, такие как выборка по экспертной оценке, или преднамеренная выборка, выборка путем случайного «блуждания», выборка по квотам и нерепрезентативная выборка, которые не удовлетворяют условиям вероятностной выборки, с изложением причин того, почему такие методы не рекомендуются для обследований домашних хозяйств.

3.2.1. Вероятностная выборка

9. Вероятностная выборка в контексте обследования домашних хозяйств включает средства, с помощью которых элементы обследуемой совокупности — территориальные единицы, домашние хозяйства и отдельные лица — отбираются для включения в обследование. Вероятностная выборка требует, чтобы: *a*) каждый элемент имел известную математическую вероятность быть отобранным, *b*) эта вероятность была выше нуля и *c*) эта вероятность была количественно рассчитана. Важно отметить, что вероятность отбора каждого элемента может быть не одинаковой, а варьироваться в соответствии с целями обследования.

10. Именно математическая природа вероятностной выборки позволяет получать по результатам обследования научно-обоснованные оценки. Более важно то, что она создает ту основу, на которой делаются умозаключения о том, что оценки выборки характеризуют всю совокупность, из которой получена данная выборка. Важнейшим побочным продуктом применения вероятностной выборки в обследованиях домашних хозяйств является то, что могут быть оценены ошибки выборки по данным, собранным на основе описываемых выборкой случаев. Ни один из указанных признаков не характерен для методов детерминированной выборки. В силу этих факторов настоятельно рекомендуется всегда использовать вероятностную выборку в обследованиях домашних хозяйств даже в том случае, если затраты на проведение обследования будут выше, чем в случае ненаучных, детерминированных методов.

3.2.1.1. Поэтапная вероятностная выборка

11. Из вышесказанного вытекает, что для выполнения обозначенных требований вероятностная выборка должна использоваться на каждом этапе процесса формирования выборки. Например, первый этап отбора, как правило, предусматривает отбор географически определенных единиц, таких как деревни. Последний этап предусматривает отбор конкретных домашних хозяйств или лиц для последующего опроса. Для формирования надлежащей выборки на этих двух этапах и на любых промежуточных этапах должны использоваться вероятностные методы. Ниже для иллюстрации приводится упрощенный пример.

Пример

Предположим, что простая случайная выборка (ПСВ) из 10 деревень отобрана из общего числа в 100 деревень в одной сельской провинции. Предположим далее, что для каждой вошедшей в выборку деревни сделан полный список домашних хозяйств. Из этих списков для проведения опроса выполнен систематический отбор одного из каждых пяти домохозяйств, вне зависимости от того, сколько домашних хозяйств вошло в списки по каждой деревне. Это — двухэтапный план вероятностной выборки, при этом на первом этапе вероятность включения в выборку составляет 10/100, а на втором этапе — 1/5. Общая вероятность отбора конкретного домохозяйства для обследования составляет 1/50, т. е. отношение 10/100, умноженное на 1/5.

12. Не отличаясь особой эффективностью, указанный в этом примере план выборки, тем не менее, иллюстрирует то, каким образом на обоих этапах формирования выборки используется вероятностная выборка. В результате этого, оценка результатов обследования может проводиться *без погрешностей* путем надлежащего применения вероятностей отбора на этапе анализа данных в процессе проведения обследования (см. изложение проблемы взвешивания результатов обследования в главе 6).

3.2.1.2. Расчет вероятности

13. Приведенный выше пример также иллюстрирует, каким образом были удовлетворены два других требования к вероятностной выборке. Во-первых, каждой деревне в данной провинции придана *отличная от нуля* вероятность включения в выборку. В противном случае, если бы одна или более деревень были исключены из рассмотрения по любой причине, в том числе по соображениям безопасности, вероятность включения в выборку таких деревень была бы равна нулю, и, следовательно, вероятностный характер выборки был бы нарушен. Домашние хозяйства в приведенном выше примере также были отобраны с отличной от нуля вероятностью. Однако, если бы некоторые из них были целенаправленно исключены, например, в связи с их недоступностью,

у них была бы равная нулю вероятность включения, и в этом случае осуществление выборки вернулось бы к детерминированному формату. В разделе 3.2.1.3 рассматриваются пути решения проблемы, возникающей, когда те или иные районы исключаются из обследования.

14. Во-вторых, вероятность включения в выборку как деревень, так и домашних хозяйств фактически может быть *рассчитана* на базе имеющейся информации. Применительно к отбору деревень были известны и размер выборки (10), и размер совокупности (100), и именно эти параметры определили вероятность, равную 10/100. Для домашних хозяйств расчет вероятности был несколько иным, поскольку до обследования нам не было известно, сколько домохозяйств должны быть отобраны в каждой из вошедших в выборку деревень. Нам было просто указано выбрать одно домохозяйство из пяти. Следовательно, если бы в деревне А было всего 100 домохозяйств, а в деревне В — 75, мы бы отобрали, соответственно, 20 и 15 из них. По-прежнему, вероятность отбора любого домохозяйства составляла бы 1/5 вне зависимости от размера совокупности и размера выборки (20/100 = 1/5, и так же для 15/75).

15. Вновь ссылаясь на приведенный выше пример, следует отметить, что вероятность отбора на втором этапе могла бы быть рассчитана путем перекрестной проверки после завершения обследования. Если известны величины m_i и M_i, где m_i и M_i представляют собой, соответственно, число домашних хозяйств в выборке и общее число домашних хозяйств в i-ой деревне, эта вероятность будет равна m_i/M_i. Всего таких вероятностей будет 10 — по одной для каждой деревни, включенной в выборку. Однако, как было отмечено, это соотношение для указанного плана выборки всегда будет равно 1/5. При этом было бы излишним делать подсчет выборки и общего числа домохозяйств с единственной целью — рассчитать вероятность для второго этапа. Тем не менее в целях *контроля качества* было бы полезно получить такие расчеты для обеспечения правильности применения доли выборки, равной 1 из 5.

3.2.1.3. Случаи, когда неправильно определена обследуемая совокупность

16. Иногда условия вероятностной выборки нарушаются по причине расплывчатости критериев определения *обследуемой совокупности*. Например, к предназначенной для обследования совокупности могут относиться все домашние хозяйства страны. Однако при составлении плана/проведении обследования из него зачастую преднамеренно исключаются определенные подгруппы населения, такие как кочевые домохозяйства, команды на судах, а также группы населения, проживающие в труднодоступных районах. В других случаях из обследуемой совокупности, предназначенной для охвата ограниченных, особых групп населения, таких как женщины, когда-либо состоявшие в браке, или молодежь в возрасте до 25 лет, по различным причинам исключаются важные по численности подгруппы. Например, из обследуемой совокупности, охватывающей молодежь в возрасте до 25 лет, могут исключаться лица, находящиеся на военной службе, или в заключении, или в иных специальных учреждениях.

17. В любом случае, когда фактическая обследуемая совокупность отличается от задуманной ранее, группа, проводящая обследование, должна уделить особое внимание более точному повторному определению обследуемой совокупности. Это представляется важным не только для разъяснения пользователям результатов обследования, но и для соответствия условиям вероятностной выборки. В указанном выше примере касательно молодежи в возрасте до 25 лет необходимо более точно описать и повторно определить обследуемую совокупность, как *молодежь в возрасте до 25 лет, не находящуюся на военной службе или в специальных учреждениях*. В противном случае сферу охвата обследования необходимо расширить путем включения в него указанных исключенных подгрупп.

18. Следовательно, важно весьма тщательно определять обследуемую совокупность, с тем чтобы она охватывала только такие единицы, которым фактически будет предоставлена *вероятность быть отобранными для* обследования. Совершенно очевидно, что в тех случаях, когда преднамеренно исключаются те или иные подгруппы, чрезвычайно важно применять вероятностные методы к фактической совокупности, составляющей инструментарий выборки. Более того, руководители обследования должны взять на себя обязанность четко разъяснить пользователям после опубликования результатов работы, какие сегменты населения включены в обследование, а какие — исключены из него.

3.2.2. Методы детерминированной выборки

19. В отличие от вероятностной выборки не существует статистической теории, которая бы лежала в основе использования детерминированной выборки. Она может оцениваться только субъективными методами. Отказ от использования вероятностных методов означает, таким образом, что оценки по результатам обследования будут иметь погрешность. Более того, масштаб таких погрешностей, а подчас и их направленность в сторону недооценки или переоценки будут неизвестны. Как указывалось выше, точность выборочных оценок, а иными словами, их среднеквадратическая ошибка, могут быть оценены с применением вероятностной выборки. Это необходимо пользователям для оценки надежности показателей обследования и для построения доверительных интервалов вокруг этих показателей. Имеющие погрешность оценки могут появляться при вероятностной выборке в определенных условиях, например в случае, когда возникает необходимость в соответствии обследуемой совокупности другим средствам контроля (более подробно данный вопрос рассматривается в главе 6).

20. Несмотря на отсутствие теоретического обоснования, детерминированная выборка зачастую используется в различных условиях и ситуациях. Практикующие ее специалисты обычно обосновывают это низкими затратами, удобством или даже сомнением группы, проводящей обследование, что «случайная» выборка может не быть достаточно репрезентативной для обследуемой совокупности. В контексте обследований домашних хозяйств мы дадим краткий обзор различных видов детерминированной выборки, в основном с помощью примеров, и укажем некоторые из причин, по которым их не следует использовать.

3.2.2.1. Выборка по экспертной оценке

21. Выборка по экспертной оценке представляет собой метод, полагающийся на «экспертов» при отборе элементов выборки. Ее сторонники утверждают, что этот метод устраняет возникающую при использовании методов рандомизации возможность формирования «плохой» или смещенной выборки, как, например, такой выборки, при которой все элементы выборки неудачно концентрируются, скажем, в северо-западном регионе.

Пример

В качестве примера выборки по экспертной оценке при обследовании домашних хозяйств можно привести группу экспертов, которые целенаправленно отобрали географические районы в качестве элементов на первом этапе отбора в плане выборки и которые обосновали свое решение тем мнением, что данные районы являются типичными или репрезентативными в том или ином смысле или контексте.

22. Основной проблемой этого вида выборки является субъективность определения того, какой набор районов является репрезентативным. Как ни парадоксально, но этот выбор в значительной

степени зависит от *выбора* самих экспертов. И наоборот, при применении вероятностной выборки эти районы будут сначала стратифицированы с использованием при необходимости любых критериев по желанию проектной группы. Следует отметить, что критерии стратификации могут даже быть *субъективными*, хотя существуют рекомендации по применению более объективных критериев (см. раздел 3.4 по вопросам стратификации). Тогда вероятностная выборка районов (отобранных одним из множества имеющихся способов) будет сформирована *из каждой страты*. Следует также отметить, что стратификация значительно снижает вероятность формирования смещенной выборки, подобной упомянутой выше. *Именно по этой причине был изобретен метод стратификации*. При стратифицированной выборке каждый район имеет известную, отличную от нуля вероятность отбора, которая является несмещенной и не подверженной субъективному мнению (даже в том случае, если сами страты будут определяться субъективно). С другой стороны, выборка по экспертной оценке не предусматривает ни механизма обеспечения отличной от нуля вероятности отбора для каждого района, ни механизма расчета вероятности отбора тех районов, которые в конечном счете включаются в выборку.

3.2.2.2. Выборка с помощью случайного «блуждания» или выборка по квотам

23. Другим широко используемым видом детерминированной выборки является так называемая процедура случайного «блуждания», осуществляемая на последнем этапе обследования домашних хозяйств. Этот метод зачастую применяется даже в том случае, если элементы выборки предыдущих этапов были отобраны с помощью надлежащих вероятностных методов. В приводимом ниже примере показан вид выборки, представляющий собой комбинацию выборки методом случайного блуждания и выборки по квотам. Последняя из них является еще одним методом детерминированной выборки, в которой регистраторам предоставляются квоты на опрос определенных категорий лиц.

Пример

В качестве иллюстрации этого метода регистраторам будут даны инструкции начать процесс опроса в некоторой произвольно выбранной географической точке, скажем, в деревне, и следовать по оговоренному маршруту для выбора подлежащих опросу домашних хозяйств. Инструкция может предусматривать выбор для опроса каждого n-го домохозяйства или проверку каждого находящегося на маршруте домохозяйства на предмет наличия в нем представителей конкретной исследуемой группы населения, например детей младше 5 лет. В последнем случае в ходе обследования опрашивается каждое подпадающее под такой критерий домохозяйство вплоть до достижения какой-либо заранее определенной квоты.

24. Применение этой методологии зачастую обосновывается необходимостью избежать больших затрат средств и времени, понесенных на предыдущем этапе составления списка всех домашних хозяйств в районе выборки — деревне, кластере или сегменте — до отбора домохозяйств, подлежащих опросу. Ее использование также обосновывается возможностью избежать неполучения ответов, поскольку регистратор продолжает двигаться дальше от не давших ответы домохозяйств до тех пор, пока не наберет достаточное для выполнения квоты число ответивших домохозяйств. Кроме того, сторонники этого метода утверждают, что он остается не смещенным, если начальная точка маршрута выбирается на случайной основе. Они также утверждают, что есть возможность надлежащим образом рассчитать вероятности выбора, поделив отобранное число домохозяйств на общее число домохозяйств в данной деревне, исходя из того, что последний показатель либо известен, либо может быть достаточно точно оценен.

25. С учетом изложенных в предыдущем пункте условий, теоретически возможно получить вероятностную выборку. Однако на практике сомнительно когда-либо в действительности достичь этой цели. Такой подход обычно не работает в силу *a*) действий регистраторов и *b*) отношения к не ответившим домохозяйствам, включая те, которые могут потенциально войти в категорию не ответивших. Как показали многочисленные исследования, когда регистраторам предоставляются функции определять выборку на местах, результатом этого будет пристрастная выборка. Например, средний размер (число лиц) вошедших в выборку домашних хозяйств обычно бывает ниже, нежели число лиц, проживающих в этих домохозяйствах[1]. Исходя из свойств человеческой натуры, регистратор будет избегать опроса домохозяйства, в котором он, с его точки зрения, может столкнуться с какими-либо сложностями. По этой причине бывает проще обойти стороной домашнее хозяйство со злой собакой или то, которое находится за глухими воротами и выглядит недоступным, и выбрать вместо этого соседнее домохозяйство, с которым подобных проблем не возникает.

26. После замены не ответивших домохозяйств на ответившие выборка становится смещенной в сторону готовых к сотрудничеству и легко достижимых домохозяйств. Совершенно ясно, что имеют место различия в характеристиках домохозяйств в зависимости от их желания и готовности к участию в обследовании. При использовании метода выборки по квотам те лица, с которыми сложно войти в контакт или которые не хотят сотрудничать, будут с большой вероятностью недопредставлены в выборке по сравнению с вариантом использования вероятностной выборки. В последнем случае от регистраторов, как правило, требуют нескольких повторных посещений домохозяйств, члены которых временно отсутствуют. Более того, применительно к обследованиям, основанным на вероятностной выборке, регистраторов обычно готовят к использованию дополнительных мер для убеждения колеблющихся домохозяйств принять участие в опросе.

3.2.2.3. Нерепрезентативная выборка

27. Нерепрезентативная выборка также широко используется в силу простоты ее формирования. Хотя такая выборка и редко применяется в обследованиях домашних хозяйств, можно представить много примеров ее использования, скажем, при проведении обследования школьников старших классов в целенаправленно сформированной выборке школ, которые известны своей доступностью и готовностью к сотрудничеству, иными словами, «удобных». Другое, модное сейчас применение — это мгновенный опрос на Интернет-сайтах, когда заходящих на такие сайты пользователей просят высказать свое мнение по различным темам. Вполне очевидно, почему такая выборка является по своей сути смещенной и не должна использоваться для выводов по населению в целом.

3.3. Определение размера выборки для обследований домашних хозяйств

28. Данный раздел носит весьма детальный характер в силу важности определения размера выборки для всех операций и затрат в ходе обследования. Размер выборки важен не только с точки зрения того, какое число домашних хозяйств будет опрошено, но также и в плане того, сколько первичных единиц выборки (ПЕВ) из географических районов вошли в выборку, сколько надо

1 В настоящее время стандартной практикой многих проводящих обследования организаций является обеспечение того, чтобы определение домохозяйств для включения в выборку проводилось в кабинетных условиях, где этот процесс легко поддается надзору со стороны контролеров. Более того, выборка должна формироваться сотрудником, который не участвовал в создании списка домохозяйств до начала формирования выборки или не знаком с фактическим положением дел на местах.

привлечь регистраторов, насколько большой объем работы будет приходиться на одного регистратора и т. д. Существует множество факторов и параметров, которые необходимо учитывать при определении размера выборки, однако все они в значительной мере касаются целей измерения в рамках конкретного обследования. Мы рассмотрим проблему определения размера выборки по следующим аспектам: основные оценки, которые требуется получить, обследуемые совокупности, число домохозяйств, которые должны войти в выборку для оценки соответствующих обследуемых совокупностей, требуемая точность и доверительная вероятность, области оценки, необходимость измерения уровня того или иного показателя или его изменения, формирование кластеров, допуск на неполучение ответов и имеющийся бюджет. Совершенно очевидно, что размер выборки — это ключевой фактор, определяющий общий план выборки.

3.3.1. Числовые значения оценок обследования

29. В обследованиях домашних хозяйств как общей направленности, так и посвященных определенной теме, такой как здравоохранение или экономическая активность, каждая оценка (часто называемая *показателем*), которую необходимо получить из обследования, требует разных размеров выборки для надежного измерения. Размер выборки зависит от величины оценки, т. е. от ее ожидаемой доли в генеральной совокупности. Например, для надежной оценки доли домашних хозяйств, имеющих доступ к безопасной воде, требуется иной размер выборки, чем для оценки доли взрослого, не работающего на данный момент населения.

30. Выражения для расчета размеров выборки базируются на вероятностных предположениях, что истинный параметр совокупности находится в интервале с данной вероятностью (доверительная вероятность). Ширина (или точность) этого интервала зависит от значения вариантности совокупности, указанной в таблице 3.2, а также от доверительного уровня и от размера выборки. Как правило, чем выше однородность совокупности или желаемая доверительная вероятность, тем шире будет такой интервал. С другой стороны, ширина интервала будет сокращаться по мере увеличения размера выборки. Примеры доверительных интервалов приводятся в пункте 22 главы 7. Указанное ниже выражение определяет доверительный интервал среднего значения $\bar{Y}$ совокупности, принимая во внимание оценку среднего значения $\hat{\bar{Y}}$ совокупности, базирующуюся на простой случайной выборке без замещения, имеющей размер n.

$$P\left[\hat{\bar{Y}} - z_{1-\alpha}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2(y)}{n}} \le \bar{Y} \le \hat{\bar{Y}} + z_{1-\alpha}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2(y)}{n}}\right] = (1-\alpha)100\%\,, \qquad (3.5)$$

где $1-\alpha$ — это доверительный коэффициент для данного интервала. Следует отметить также, что применительно к оценке соотношения p, значение $\sigma^2(y) = p(1-p)$.

31. На практике обследование само по себе может иметь только один размер выборки. Для расчета размера выборки необходимо сделать выбор среди многих оценок, которые должны быть получены в ходе того или иного обследования. Например, если ключевой оценкой является уровень безработицы, расчет размера выборки должен базироваться именно на нем[2]. При наличии

2 Несколько парадоксальным является тот факт, что формула расчета размера выборки требует знания приблизительной величины оценки, подлежащей измерению. Эту величину, однако, можно «предположить», используя различные методы, например, путем использования данных переписи или аналогичного обследования, данных по соседней стране, экспериментального обследования и т. д.

многих ключевых показателей иногда используется методика, по которой рассчитывается размер выборки, необходимый для каждого из них, а затем используется тот, который дает самую крупную по размеру выборку. Как правило, таким оказывается показатель, для которого базовая совокупность представляет собой наименьшую «подгруппу обследуемой совокупности» с точки зрения ее доли в генеральной совокупности. Необходимо, конечно, принимать во внимание желаемый уровень точности (см. ниже). Когда размер выборки базируется на такой оценке, каждая из прочих ключевых оценок будет измеряться с тем же или более высоким уровнем надежности.

32. В качестве альтернативы размер выборки может базироваться на сравнительно малой доле обследуемой совокупности вместо выделения какого-то определенного показателя. Это, по всей видимости, является наилучшим подходом для общих обследований домашних хозяйств, в которых упор делается на нескольких не связанных друг с другом темах, и в этом случае может оказаться непрактичным или нецелесообразным базирование размера выборки на том или ином показателе, относящемся к какой-либо одной теме. Таким образом, руководители обследования могут принять решение при определении размера выборки исходить из возможности надежного измерения характеристик для 5 процентов (или 10 процентов) совокупности, при этом конкретный выбор будет зависеть от бюджетных соображений.

3.3.2. Обследуемая совокупность

33. Размер выборки также зависит от изучаемой совокупности, которая будет охватываться обследованием. Как и в случае с показателями, в обследованиях домашних хозяйств часто присутствует несколько изучаемых групп населения. Например, обследование в сфере здравоохранения может быть нацелено на *a*) домашние хозяйства для оценки их доступа к безопасной воде и санитарно-техническим системам, охватывая в то же время, *b*) всех лиц для оценки хронических и тяжелых нарушений здоровья, *c*) женщин в возрасте 14–49 лет для выявления показателей их репродуктивного здоровья и *d*) детей в возрасте до 5 лет для антропометрических измерений роста и веса.

34. Следовательно, необходимо рассмотреть размер выборки для каждой из этих исследуемых групп населения. Как указано выше, обследования домашних хозяйств зачастую включают несколько изучаемых совокупностей, каждая из которых представляет одинаковый интерес, с точки зрения целей обследования, в плане измерения тех или иных показателей. И вновь здесь имеет смысл сконцентрироваться на наименьшей совокупности для определения размера выборки. Например, если важнейшей целевой группой обследования являются дети младше 5 лет, размер выборки должен базироваться именно на этой группе. Применяя подход, описанный в пункте 32, руководящая группа по проведению обследования может принять решение по расчету размера выборки для оценки характеристики в отношении 10 процентов детей в возрасте до 5 лет. Полученный в результате размер выборки будет значительно больше, чем необходимо для целевой группы, состоящей из всех лиц, проживающих во всех домохозяйствах.

3.3.3. Точность и статистическая достоверность

35. Выше было приведено соображение о том, что оценки, особенно применительно к ключевым показателям, должны быть *надежными*. Определение размера выборки в значительной мере зависит от желаемого уровня точности показателей. Чем более точными или надежными должны быть результаты обследования, тем больше (на порядки величины) должен быть размер выборки. Например, удвоение требования надежности может обусловить необходимость увеличения размера выборки *в четыре раза*. Безусловно, руководители обследования должны быть осведомлены

о том влиянии, которое оказывают чрезмерно строгие требования к точности на размер выборки и, следовательно, на уровень затрат при проведении обследования. И наоборот, они должны тщательно избегать слишком малых размеров выборки, в результате чего основные показатели будут слишком ненадежными для информативного анализа или эффективного планирования.

36. Аналогичным образом, размер выборки увеличивается по мере роста желаемого уровня статистической достоверности для сохранения определенной точности. В качестве стандарта практически повсеместно принимается 95-процентный уровень доверительной вероятности, и соответствующим образом рассчитывается необходимый для достижения этого уровня размер выборки (см. пункт 30, выше).

37. Принимая во внимание те или иные показатели, методикой многих хорошо спланированных обследований является использование в качестве требуемого уровня точности предельного уровня *относительной* ошибки в 10 процентов при 95-процентном уровне доверительной вероятности по подлежащим измерению ключевым показателям, и это по сути означает, что среднеквадратическая ошибка того или иного ключевого показателя не должна превышать 5 процентов от самой оценки. Это рассчитывается, как ($2 \times 0{,}05x$, где x — это оценка в ходе обследования). Например, если прогнозируемая доля населения в составе рабочей силы составляет 65 процентов, ее среднеквадратическая ошибка не должна превышать 3,25 процентных пункта, т. е. значение 0,65, умноженное на 0,05. Значение 0,0325, помноженное на два, или 0,065 дает предельный уровень относительной ошибки при 95-процентной доверительной вероятности. Например, как указано в пункте 30, выше, мы имеем:

$$\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2(y)}{n}} = 0{,}05 \times \hat{\bar{Y}}. \qquad (3.6)$$

38. Размер выборки, необходимый для достижения критерия, равного 10-процентному предельному уровню относительной ошибки, таким образом, представляет собой одну четвертую единицы, где предельный уровень относительной ошибки установлен на уровне 5 процентов. Предельный уровень относительной ошибки в 20 процентов обычно считается максимально допустимым для важных показателей (хотя мы не рекомендуем такой уровень). Это обусловлено тем, что доверительный интервал применительно к оценкам с более высокими допусками на ошибку слишком широк для получения содержательных результатов для большинства аналитических или стратегических целей. Как правило, при наличии соответствующего бюджета, мы рекомендуем предельный уровень относительной ошибки в 5–10 процентов в отношении основных показателей.

3.3.4. Группы анализа: области обследования

39. Еще одним существенным фактором, оказывающим большое влияние на размер выборки, является число областей обследования. Области обследования, как правило, определяются как аналитические подгруппы, для которых необходимы *равнозначно* надежные данные. Размер выборки увеличивается приблизительно[3] на множитель, равный числу искомых областей обследования. Это, однако, является справедливым, если каждая из областей обследования демонстрирует аналогичный уровень изменчивости (для дополнительного разъяснения см. сноску 3).

3 Это справедливо для любых случаев, когда требуется одинаковый уровень надежности для каждой из областей обследования.

Это происходит потому, что размер выборки для данного уровня точности не зависит от размера самой совокупности, за исключением случаев, когда размер выборки составляет существенную процентную долю совокупности — например 5 и более процентов (что редко встречается в случае обследований домашних хозяйств). Следовательно, размер выборки, требуемый для одной провинции (если обследование ограничено только одной провинцией), будет таким же, какой необходим для всей страны в целом. Это чрезвычайно важный момент, который зачастую неправильно понимается специалистами-практиками по проведению обследований, которые ошибочно полагают, что чем крупнее совокупность, тем больше должен быть размер выборки.

40. Таким образом, когда требуются лишь данные на национальном уровне, существует единственная область обследования, и соответствующим образом рассчитанный размер выборки применяется для выборки по всей стране в целом. Однако, если принято решение о том, что результаты одинаковой надежности необходимо получить по городским и сельским районам по отдельности, то тогда требуемые размеры выборки будут рассчитаны для каждой области обследования в целях получения надежных результатов. Как правило, размер выборки для каждой из соответствующих областей обследования должен рассчитываться таким образом, что если будет $D_1, D_2, \ldots \ldots D_k$ областей обследования, то обязательно будет $n_1, n_2, \ldots \ldots n_k$ размеров выборки, которые будут зависеть от изменчивости соответствующих характеристик каждой из областей обследования, а также от установленных уровней доверительной вероятности и точности. Таким образом, общий размер выборки для всего обследования будет равняться $n = n_1 + n_2 + \ldots\ldots + n_k$.

3.3.4.1. Дополнительная выборка для оценок по областям обследования

41. Важным следствием требования равной надежности для областей обследования является необходимость использования непропорциональных долей выборки. Таким образом, когда распределение отклоняется от соотношения 50-50, что весьма вероятно для городских и сельских областей обследования, по всей видимости, в большинстве стран возникнет необходимость в применении преднамеренной дополнительной выборки, например, в городском секторе для достижения равной надежности. При этом следует подчеркнуть, что дополнительная выборка в той или иной области изучения в ходе общенационального обследования в основном диктуется необходимостью получения результатов с определенным доверительным уровнем.

42. Важно отметить два последствия применения преднамеренной дополнительной выборки в отношении подгрупп, как в областях, так и в стратах обследования. Во-первых, она требует применения в рамках обследования компенсирующих весовых коэффициентов для получения оценок на национальном уровне. Во-вторых, и что более важно, национальные оценки будут несколько менее надежными, чем они были бы в случае пропорционального распределения выборки среди подгрупп.

3.3.4.2. Отбор областей обследования

43. Подразделения территориальных единиц безусловно важны, и всегда существует соблазн рассматривать их в качестве областей обследования для целей получения соответствующих оценок. Например, при проведении общенационального обследования пользователи данных во властных структурах хотят получить данные не только по каждому крупному региону, но и зачастую по каждой провинции. Совершенно очевидно, что необходимо внимательно подойти к вопросу о числе областей обследования и обоснованно выбрать вид групп оцениваемых показателей, характеризующих данные области. Одной из приемлемых здесь стратегий является принятие решения о том, какие группы оценок, вне зависимости от их важности, требуют *равнозначной* надежности измерений в рамках обследования. Вместо этого группы оценок будут при анализе

рассматриваться в качестве основных составных *категорий* таблиц, а не областей обследования. В этом случае размеры выборки для каждой такой категории будут меньше, чем если бы они рассматривались в качестве областей обследования; вследствие этого более низкой будет также и их надежность. При этом, однако, следует отметить, что дополнительная выборка в какой-либо определенной области обследования может быть обусловлена необходимостью получения оценок по данной области с определенными уровнями доверительной вероятности и точности вне зависимости от соответствующих уровней, установленных на национальном уровне.

Пример

Следующий пример показывает, каким образом будет сделана выборка и каково будет ее влияние на надежность оценок, если городские и сельские районы будут рассматриваться как группы категорий таблицы, а не как области обследования. Предположим, что распределение населения составляет: 60 процентов — сельское население и 40 процентов — городское. Если для удовлетворения оговоренного требования точности расчетный размер выборки был определен на уровне 8 000 домашних хозяйств, то в случае, если городские и сельские районы устанавливаются как отдельные области обследования, 16 000 из них должны войти в такую выборку — по 8 000 домохозяйств в каждом секторе. Вместо этого, рассматривая их в качестве групп в таблице, будет сформирована национальная выборка в количестве 8 000 домохозяйств в соответствующей пропорции из городских и сельских районов, что дает, соответственно, 4 800 и 3 200 домохозяйств. Предположим далее, что ожидаемая среднеквадратическая ошибка для 10-процентной характеристики, базирующейся на выборке из 8 000 единиц, составит 0,7 процентных пункта. Эта среднеквадратическая ошибка применяется в отношении оценки на национальном уровне (или в отношении городских и сельских районов по отдельности, если по каждой области обследования в выборку включены по 8 000 домохозяйств). Для национальной выборки в 8 000 домохозяйств, пропорционально сформированной из городских и сельских районов, соответствующая среднеквадратическая ошибка для сельских районов составит примерно 0,9 процентных пункта, при ее расчете как произведения квадратного корня отношения размеров выборки на среднеквадратическую ошибку национальной оценки, или

$$\left(\left[\sqrt{8\,000/4\,800}\right]\times 0{,}7\right).$$

Для городских районов среднеквадратическая ошибка составит примерно 1,1 процентных пункта, или

$$\left(\left[\sqrt{4\,800/3\,200}\right]\times 0{,}7\right).$$

Другим способом этот эффект можно оценить, исходя из того факта, что среднеквадратические ошибки для всех сельских оценок будут примерно на 29 процентов выше

$$\left(\sqrt{8\,000/4\,800}\right),$$

чем для национальных оценок; для городских районов они будут примерно на 58 процентов выше

$$\left(\sqrt{8\,000/3\,200}\right).$$

44. Следует отметить, что последнее суждение в вышеуказанном примере применимо вне зависимости от того, какая среднеквадратическая ошибка установлена на национальном уровне. Иными словами, оно применимо к любой оценке, включенной в таблицу обследования. Следовательно, появляется возможность анализа влияния на надежность данных еще до формирования выборки для различных подгрупп, которые могут рассматриваться в качестве областей обследования. Используя такой подход, группа по проведению обследования будет располагать информацией, которая поможет ей принять решение о том, следует ли рассматривать потенциальные области обследования в качестве групп таблицы. Как следует из вышесказанного, это означает, что надлежит использовать пропорциональное, а не равное распределение выборки. Например, если общенациональное обследование планируется в стране, имеющей только 20 процентов городского населения, размер выборки для городских районов будет составлять лишь 20 процентов от общего размера выборки. Следовательно, ошибка выборки для городских оценок будет в два раза (квадратный корень из 0,8n/0,2n) превышать ошибку выборки для сельских оценок и примерно в два с четвертью раза (квадратный корень из n/0,2n) превышать ошибку выборки для национальных оценок. В этом случае руководители обследования могут принять решение о необходимости дополнительной выборки в городском секторе[4], создавая, таким образом, отдельные городские и сельские области обследования.

45. Аналогичным образом можно проанализировать взаимосвязь между среднеквадратическими ошибками и областями обследования по сравнению с группами таблицы для помощи в принятии решения о том, стоит ли использовать регионы или другие субнациональные административные единицы в качестве областей обследования, и если да — то какое их количество следует использовать. При равных размерах выборки, необходимых для областей обследования, использование 10 регионов потребует выборку в 10 раз больше размера национальной выборки, однако этот размер сократится вдвое, если будет признано, что лишь пять регионов достаточно отобрать для выполнения стратегических целей. Аналогичным образом, если вместо этого регионы будут рассматриваться как группы таблицы, национальная выборка будет распределена между ними пропорционально. В этом случае *средний* регион будет иметь среднеквадратическую ошибку примерно в 3,2 раза выше, чем национальные оценки в случае 10 регионов, и только в два раза выше — в случае пяти регионов.

3.3.5. Влияние гнездовой группировки

46. В данном разделе рассматривается вопрос о том, каким образом формирование кластеров влияет на размер выборки (более подробно проблема гнездовой выборки рассматривается в разделе 3.5). Та степень, в которой выборка обследования домашних хозяйств *сгруппирована по кластерам*, влияет на надежность или точность оценок, а следовательно, и на размер выборки. Влияние гнездовой группировки при обследовании домашних хозяйств возникает из-за *a*) предпоследних единиц выборки, обычно называемых «кластерами», в качестве которых могут выступать деревни или городские микрорайоны, *b*) вошедших в выборку домохозяйств, *c*) размера и/или изменчивости кластеров и *d*) метода включения в выборку домохозяйств в рамках отдельных кластеров. Влияние гнездовой группировки, а также стратификации может быть численно измерено с помощью эффекта схемы или *deff*, который является выражением того, насколько возрастает вариантность выборки (квадрат среднеквадратической ошибки) для стратифицированной гнез-

4 Такое решение может быть принято, например, если предполагаемые относительные среднеквадратические ошибки для (любого из) ключевых городских показателей будут выше, к примеру, на 7,5 процента (95-процентная доверительная вероятность составит 15 процентов, что в настоящем руководстве рассматривается как максимально допустимый уровень).

довой выборки по сравнению с простой случайной выборкой такого же размера. Стратификация склонна *снижать* вариантность выборки, однако, лишь незначительно. И наоборот, гнездовая группировка значительно увеличивает вариантность. Таким образом, *deff* в основном показывает, в каком объеме гнездовая группировка присутствует в выборке обследования.

47. Эффективный план выборки требует использования кластеров для снижения затрат при одновременном сохранении эффекта схемы на низком по возможности уровне для того, чтобы результаты обследования были пригодными к использованию и надежными. К сожалению, величина *deff* не известна до проведения обследования, и может быть оценена только по его итогам из самих данных. В случаях, когда проводились предыдущие обследования или аналогичные обследования в других странах, величины *deff* из таких обследований можно использовать в качестве косвенных показателей в формуле расчета для оценки размера выборки.

48. Для сохранения эффекта схемы на низком по возможности уровне план выборки должен следовать следующим общим принципам (см. также краткие рекомендации в конце данной главы):

a) использование максимального практически возможного числа кластеров;

b) использование наименьшего практически возможного размера кластеров с точки зрения числа домохозяйств;

c) использование постоянного, а не переменного размера кластеров;

d) формирование на последнем этапе систематической выборки домохозяйств, которые географически разбросаны, а не сегмента территориально прилегающих друг к другу домохозяйств.

49. Таким образом, для выборки в 12 000 домашних хозяйств предпочтительно сформировать 600 кластеров по 20 домохозяйств в каждом, чем 400 кластеров по 30 домохозяйств в каждом. В первом случае эффект схемы выборки будет значительно меньше. Более того, величина *deff* снижается, если из всех домохозяйств одного кластера домохозяйства отбираются систематически, а не из территориально прилегающих друг к другу подсегментов. При соблюдении этих эмпирических правил эффект схемы будет оставаться в разумно низких пределах.

3.3.6. Поправка на прогнозируемое неполучение ответов при определении размера выборки

50. Общепринятой практикой при проведении обследований является увеличение размера выборки на величину, равную прогнозируемой доле неполучения ответов. Это обеспечивает то, что фактическое число опросов, проведенных в рамках обследования, будет приблизительно соответствовать целевому размеру выборки.

51. Показатели неполучения ответов в обследованиях широко варьируются по различным странам и видам обследований. В приведенных ниже расчетах мы исходим из предполагаемой доли неполучения ответов на уровне 10 процентов. Безусловно, любая страна должна использовать цифру, которая наиболее точно отражает опыт последних национальных обследований.

3.3.7. Размер выборки для эталонных выборок

52. Эталонные выборки подробно рассматриваются в главе 4. В данном разделе упор делается на размере выборки для плана эталонной выборки. Вкратце, эталонная выборка — это крупная выборка ПЕВ для тех стран, которые осуществляют крупномасштабные и непрерывные комплекс-

ные программы обследований. Такая крупная выборка предназначена для предоставления достаточного числа «зарезервированных» типов выборки в целях поддержки проведения многочисленных обследований в течение нескольких лет без необходимости многократного опроса одних и тех же респондентов.

53. В условиях, когда многочисленные обследования и, следовательно, многочисленные темы охвачены эталонной выборкой, безусловно, существуют и многочисленные изучаемые совокупности, которые должны быть обследованы, и ключевые оценки, которые должны быть получены. В этой связи большинство стран формируют размер выборки, исходя из двух соображений. Во-первых, и это само собой разумеется, присутствуют бюджетные соображения. Во-вторых, это — предполагаемые и охватываемые эталонной выборкой размеры выборки отдельных обследований, которые могут использоваться в течение определенного периода времени, часто вплоть до 10 лет между переписями населения. Следовательно, возможные размеры выборки для эталонных выборок доходят до весьма крупных размеров, достигая 50 000 домашних хозяйств и более. Тщательно отработаны планы применения всего архива домашних хозяйств.

Пример

Предположим, что эталонная выборка страны А состоит из 50 000 домашних хозяйств. Эталонную выборку предполагается использовать в трех уже запланированных обследованиях, а также, возможно, в двух других, которые еще не запланированы. Одно из обследований касается вопроса доходов и расходов домохозяйств и должно повторяться три раза в течение десяти лет — в 1-й, 5-й и 8-й годы. В рамках каждого из трех этапов этого обследования запланирован опрос 8000 домохозяйств. На 5-м году, однако, предполагается заменить в выборке 4 000 домохозяйств, т. е. половину из 8 000 домохозяйств, опрошенных в 1-й год. Аналогичным образом, на 8-й год будут заменены на новые оставшиеся от 1-го года 4 000 домохозяйств. Таким образом, для обследования доходов и расходов будет использовано 16 000 домохозяйств. Второе запланированное обследование проводится по теме здравоохранения, в котором, как ожидается, будет использовано около 10 000 домохозяйств, а в третьем обследовании рабочей силы будет использовано примерно 12 000 домохозяйств. Суммарно 38 000 домашних хозяйств будет зарезервировано для этих трех обследований. Соответственно, 12 000 домохозяйств все еще остаются неохваченными и могут быть при необходимости использованы для других обследований.

3.3.8. Оценка изменений или уровня показателей

54. Основной целью измерений, которые проводятся в рамках обследований на периодической основе, является оценка изменений, происшедших в период между обследованиями. По статистической терминологии оценка, полученная при первом обследовании, представляет *уровень* того или иного показателя, в то время как разница между этим уровнем и оценкой уровня, полученной во втором обследовании, является расчетным *изменением*. Как правило, для оценки изменения в целях получения надежных выводов требуется значительно более крупный размер выборки, чем необходимый для оценки одного лишь уровня. Этот особенно справедливо при измерении небольших изменений. При этом, однако, существует ряд методов формирования выборки, предназначенных для сокращения размера выборки (и, следовательно, затрат на обследование) при оценке изменений (см. раздел 3.9.2).

3.3.9. Бюджет обследования

55. Само собой разумеется, что при определении надлежащего размера выборки для обследования домашних хозяйств нельзя игнорировать бюджет обследования. Хотя бюджет не является числовым параметром в математическом расчете размера выборки, он играет заметную роль на практическом уровне.

56. Первоначальными расчетами размера выборки занимаются статистики, которые учитывают каждый параметр, рассматриваемый в данной главе. При этом довольно часто такой размер оказывается больше, чем может обеспечить бюджет обследования. Когда это происходит, группа по проведению обследования должна либо просить дополнительное финансирование обследования, либо вносить изменения в свои цели измерений, снижая требования к их точности или число областей обследования.

57. В функции технического специалиста по формированию выборки входит оказание помощи в ходе обсуждения зависимости «затраты–точность». Такой специалист должен разъяснить баланс соображений, возникающих в связи с ограничением числа областей обследования (меньше практических выгод для пользователей данных) или снижением требований к точности измерений (снижение надежности ключевых показателей) во всех случаях, когда возникает необходимость сокращения надлежащего размера выборки в связи с бюджетными возможностями. Такое обсуждение должно проводиться с применением примеров по точности оценок и областям обследования, приведенных выше. Давая рекомендации группе по проведению обследования, специалист по выборке должен внимательно учитывать и тот факт, что число кластеров также является ключевым определяющим фактором расходов на обследование (этот вопрос дополнительно рассматривается в разделе 3.5.5).

3.3.10. Расчет размера выборки

58. В данном разделе мы представляем формулу расчета размера выборки, принимая во внимание рассмотренные выше параметры. В связи с тем, что основной упор мы делаем на обследованиях домашних хозяйств, размер выборки рассчитывается в плане численности подлежащих отбору домашних хозяйств. Приведен также ряд примеров.

59. Как правило, когда включено соотношение p, формула расчета[5] размера выборки, n_h, является следующей:

$$n_h = (z^2)(1-r)\ (f)\ (k)/(r)(p)\ (\tilde{n})\ (e^2), \qquad (3.7)$$

где n_h — это искомый параметр размера выборки в плане численности подлежащих отбору домашних хозяйств; z — статистическая величина, определяющая желаемую доверительную вероятность; r — величина ключевого показателя, подлежащего измерению в рамках обследования; f — эффект схемы выборки, *deff*, принимаемый равным 2,0 (значение по умолчанию); k — множитель, необходимый для учета предполагаемой доли неполучения ответов; p — доля генеральной

5 Формула для размера выборки может также содержать некий множитель, так называемый поправочный множитель на конечность выборки, который необходимо учитывать в случаях, когда результирующий расчетный размер выборки составляет значительную долю размера совокупности. Однако это условие редко возникает применительно к крупномасштабным обследованиям домашних хозяйств тех типов, которые рассматриваются в данном руководстве. В нашем случае поправочный множитель на конечность выборки принимается за 1,0 и, следовательно, не учитывается в формуле 3.5.

совокупности, вошедшая в обследуемую совокупность и на которой базируется параметр r; $\tilde{n}$— средний размер домашнего хозяйства (число лиц, проживающих в домохозяйстве); и e — допустимый предел ошибки, к которому следует стремиться. Ниже приводятся рекомендуемые значения некоторых параметров.

60. Предполагаемая к использованию статистическая величина z должна составлять 1,96 для 95-процентного уровня доверительной вероятности (по сравнению, например, со значением 1,645 для 90-процентного уровня). Первая величина обычно считается стандартом для придания желаемого уровня доверительной вероятности при оценке допустимого предела ошибки в обследованиях домашних хозяйств. Принимаемая по умолчанию величина эффекта схемы выборки обычно устанавливается на уровне 2,0, если не существует дополнительных эмпирических данных прошлых или аналогичных обследований, доказывающих иную величину. Поправочный множитель на неполучение ответов, k, должен выбираться таким образом, чтобы отражать собственный опыт конкретной страны в части неполучения ответов в обследованиях — обычно менее 10 процентов для развивающихся стран. Таким образом, осторожным подходом к множителю k будет выбор его значения на уровне 1,1. Параметр p обычно может быть рассчитан, исходя из результатов последней переписи. Параметр $\tilde{n}$ зачастую составляет около 6,0 в большинстве развивающихся стран, однако, его необходимую для использования в формуле точную величину обычно можно получить из последней переписи. Что касается допустимого предела ошибки, e, рекомендуется, чтобы уровень точности устанавливался равным 10 процентам от r; следовательно, $e = 0{,}10r$. Меньший размер выборки может формироваться с менее строгим допустимым пределом ошибки, $e = 0{,}15r$, однако в этом случае результаты обследования безусловно будут менее надежными. Подстановка в формулу выбранных величин дает следующий результат:

$$n_h = (3{,}84)\ (1-r)\ (1{,}2)\ (1{,}1)/(r)\ (p)\ (6)\ (0{,}01). \tag{3.8}$$

Уравнение (3.2) сокращает формулу до следующего вида:

$$n_h = (84{,}5)\ (1-r)/(r)\ (p). \tag{3.9}$$

61. Сокращенный вариант может использоваться в случаях, когда вместо более точных показателей, полученных на основании опыта конкретной страны, применяются *все* рекомендуемые по умолчанию значения указанных выше параметров.

Пример

В стране B принято решение, что основным искомым показателем обследования будет уровень безработицы, который предполагается равным примерно 10 процентам от численности трудоспособного гражданского населения. Трудоспособное гражданское население определяется как население в возрасте 14 лет и старше и составляет примерно 65 процентов от общей численности населения страны. В этом случае $r = 0{,}1$ и $p = 0{,}65$. Предположим, что, как рекомендовано выше, мы хотим оценить уровень безработицы с 10-процентным допустимым пределом относительной ошибки при 95-процентном уровне доверительной вероятности; в таком случае $e = 0{,}10r$ (иными словами, среднеквадратическая ошибка составляет 0,01). Далее мы принимаем рекомендуемые значения ожидаемой доли неполучения ответов, эффекта схемы и среднего размера домашнего хозяйства. Тогда мы имеем право применить формулу

(3.9), которая в итоге дает 1 170 домашних хозяйств [(84,5×0,9)/(0,1×0,65)]. Это — сравнительно небольшой размер выборки, прежде всего в силу того, что базовая совокупность составляет большую долю от генеральной совокупности, а именно 65 процентов. Напомним, что искомый размер выборки рассчитывается для одной области обследования — в данном случае на национальном уровне. Если цели измерения предусматривают получение в равной степени надежных данных для городских и сельских районов, тогда размер выборки необходимо удвоить, исходя из того, что все параметры формул (3.8) и (3.9) применяются как для городских, так и для сельских районов. Чем больше они различаются (например, средний размер городских домохозяйств может отличаться от сельских домохозяйств, так же как могут отличаться ожидаемые доли неполучения ответов для городских и сельских районов), тем более точные величины необходимо использовать для отдельного расчета размеров выборки для городских и сельских районов. Результаты этих расчетов, безусловно, будут разными.

62. Приведенный ниже пример охватывает меньшую по размеру базовую совокупность — дети в возрасте до 5 лет.

Пример

В стране C в качестве основного показателя обследования определен уровень смертности среди детей в возрасте до 5 лет, который предполагается равным примерно 5 процентным пунктам. В этом случае $r = 0{,}05$, а p оценивается примерно в 0,15 или 0,03×5. И вновь мы хотим оценить уровень смертности с 10-процентным допустимым пределом относительной ошибки: тогда $e = 0{,}10r$ (или среднеквадратическая ошибка на уровне 0,005). Величины ожидаемой доли неполучения ответов, эффекта схемы и среднего размера домашнего хозяйства снова сохраняются на рекомендуемом нами уровне. Формула (3.9) дает около 10 704 домохозяйств (84,5×0,95)/(0,05×0,15), т. е. значительно более крупный размер выборки, чем в предыдущем примере. И опять основная причина этого связана с размером базовой совокупности, иными словами детьми в возрасте до 5 лет, численность которых составляет лишь 15 процентов от генеральной совокупности. Оцениваемый размер параметра r также является небольшим, и этот факт в комбинации с небольшой величиной p вынуждают сформировать большой размер выборки.

63. Последний пример касается того случая, когда генеральная совокупность является обследуемой совокупностью. В этом случае $p = 1$ и может не учитываться; тем не менее здесь также можно применить формулы (3.8) и (3.9) при использовании рекомендованных величин указанных выше параметров.

Пример

В стране D в качестве основного показателя обследования определена доля лиц в общем народонаселении, у которых в течение предшествующей недели возникли какие-либо острые заболевания. Эта доля оценивается на уровне между 5 и 10 процентами, и в этом случае будет использована меньшая величина, поскольку она даст более крупный размер выборки (консервативный подход). В этом случае $r = 0{,}05$, а p, естественно, равняется 1,0. Вновь мы хотим оценить уровень острых заболеваний с 10-процентным допустимым пределом относительной ошибки: $e = 0{,}10r$[6] (или среднеквадратическая ошибка на уровне 0,005), а величины ожидаемой доли неполучения ответов, эффекта схемы и среднего размера домашнего хозяйства опять сохраняются на рекомендуемом нами уровне. Формула (3.9) в итоге дает чуть более 1 600 домохозяйств (84,5×0,95)/(0,05).

6 Поскольку величина r в данном случае применяется ко всей совокупности, она равняется p, поэтому e равняется $0{,}10p$.

64. Как указывалось выше, размер выборки для обследования в конечном счете может определяться путем расчета размеров выборки для нескольких ключевых показателей и основывая выбор на том показателе, который дает наибольший размер выборки. Кроме того, до принятия окончательного решения необходимо также рассмотреть такие аспекты, как число областей обследования, а также бюджет обследования.

65. В тех странах, в которых не соблюдается одно или более из указанных выше допущений, в формуле (3.7) могут быть сделаны простые замены для получения более точных величин размера выборки. Например, средний размер домохозяйства может быть больше или меньше 6,0; доля неполучения ответов может прогнозироваться на уровне около 5 вместо 10 процентов, а величину p для какой-либо конкретной страны, как правило, можно более точно рассчитать, используя результаты переписи.

66. Рекомендуется, однако, не вносить никаких изменений в величину статистического показателя z на уровне 1,96, который является общепризнанным стандартом. Из практических соображений также следует оставить на уровне 2,0 величину эффекта схемы, f, если только, как уже указывалось, результаты последней переписи из другого источника не дают иных данных. Рекомендуется также, чтобы величина e определялась как $0{,}10r$, кроме тех случаев, когда имеющийся бюджет не может обеспечить расчетный размер выборки. В таком случае величина e может быть увеличена до $0{,}12r$ или до $0{,}15r$. При этом такие увеличения в допустимом пределе ошибки дадут в итоге гораздо более высокие ошибки выборки.

3.4. Стратификация

67. При планировании обследования домашних хозяйств широко применяемым методом является стратификация предполагаемой для обследования совокупности еще до формирования выборки. Она служит для целей классификации совокупности в подсовокупностях — стратах — на основе дополнительной информации, которая известна в отношении генеральной совокупности. Затем вне зависимости от каждой страты отбираются элементы выборки таким способом, который соответствует целям измерения в ходе обследования.

3.4.1. Стратификация и распределение выборки

68. В стратифицированной выборке размеры выборки в пределах каждой страты контролируются техническим специалистом по выборке, а не формируются путем случайного определения в процессе выборки. Разделенная на страты совокупность может иметь точно n_s единиц, отобранных из каждой страты, где n_s — это желаемое число единиц выборки в страте s. И наоборот, нестратифицированная выборка в итоге даст размер выборки для подсовокупности в страте s, который будет несколько отличаться от n_s.

Пример

Предположим, что план выборки обследования должен состоять из двух страт — городской и сельской. Из переписи населения имеется информация для разделения всех административно-территориальных единиц на городские или сельские, что позволяет стратифицировать население по данному критерию. Решено сформировать пропорциональную (в отличие от непропорциональной) выборку в каждой страте, поскольку население распределено в соотношении: 60 процентов — сельское население и 40 процентов — городское население. Если размер выборки составляет 5 000 домохозяйств, независимое формирование выборки по стратам обеспечит, что 3 000 из них будут сельскими и 2 000 — городскими. Если

бы выборка формировалась произвольно без первоначальной стратификации, распределение домохозяйств в выборке отличалось бы от соотношения 3 000–2 000, хотя такое распределение и было бы ожидаемым. Нестратифицированная выборка может для неудачного случая дать выборку в соотношении, скажем, 3 200 сельских домохозяйств и 1 800 городских.

69. Таким образом, одной из причин стратификации является снижение шанса на неудачный вариант выборки и на наличие непропорционально большого (или небольшого) числа единиц выборки, отобранных из подсовокупности, которая считается показательной для анализа. Стратификация осуществляется для обеспечения надлежащей репрезентативности важных групп подсовокупности без внесения погрешностей в проводимый отбор. Важно, однако, отметить, что надлежащая репрезентативность не подразумевает пропорциональной выборки. Во многих случаях одна или более страт могут также являться областями оценки (как обсуждалось выше). В этом случае может возникнуть необходимость формирования выборки равного размера в используемых стратах, получая, таким образом, непропорциональную выборку по стратам. Следовательно, как пропорциональное, так и непропорциональное *распределение* единиц выборки среди различных страт является вполне допустимой особенностью стратифицированной выборки, и выбор зависит от целей измерения в рамках обследования.

70. Как подразумевалось в предыдущем абзаце, стратификация может также служить средством для неявного распределения выборки, что является более простым и практичным методом, чем оптимальное распределение[7]. Иными словами, при пропорциональной выборке по отдельным стратам нет необходимости заранее рассчитывать число единиц выборки, которые должны быть распределены по каждой страте.

Пример

Предположим, что целью плана выборки является обеспечение точно пропорционального распределения общего размера выборки по каждой из 10 имеющихся в стране провинций. Если, скажем, в провинции А проживает 12 процентов населения страны, тогда в этой провинции должны быть отобраны 12 процентов кластеров выборки при условии, что ожидаемый размер выборки является постоянным. Предположим далее, что на национальном уровне необходимо выбрать всего 400 кластеров. Часто используемый во многих странах метод состоит в *присвоении* 48 (0,12 × 400) кластеров провинции А. Однако при надлежащей стратификации такая процедура становится ненужной. Вместо этого каждая провинция должна рассматриваться в качестве отдельной страты в процессе формирования выборки. Тогда применение систематической выборки с вероятностью, пропорциональной размеру (см. таблицу 3.1), с единым интервалом выборки автоматически приведет к искомым 48 кластерам в провинции А. Этот вид стратификации, а также возможности ее использования для упрощения схем распределения более подробно рассматриваются в разделе 3.4.3.

3.4.2. Правила стратификации

71. Существуют два базовых правила, которые применяются при стратификации какой-либо совокупности. Одно из правил необходимо соблюдать всегда. Обычно следует соблюдать и другое правило, хотя его несоблюдение не приведет к серьезному ущербу для плана выборки. Требуемое

7 Оптимальное распределение означает распределение на основе стоимостных функций и различных величин вариантности в рамках страт (мер неоднородности). Этот метод не рассматривается в данном руководстве, поскольку он редко используется на практике в развивающихся странах. Это может быть обусловлено отсутствием точных цифр по расходам на операции обследования. Читатель может найти подробную информацию по оптимальному распределению во многих ссылках, указанных в конце данной главы.

правило заключается в том, что по крайней мере одна единица выборки должна отбираться из каждой создаваемой страты. Страты являются по сути независимыми и взаимоисключающими подгруппами совокупности: каждый элемент совокупности должен присутствовать в одной и только в одной страте. В силу такой особенности каждая страта *должна* участвовать в выборке, с тем чтобы в выборку могла войти вся совокупность и была рассчитана несмещенная оценка совокупности. Поскольку каждая страта может теоретически рассматриваться независимо в плане выборки, нет необходимости создавать страты, используя объективные критерии; при желании могут применяться и субъективные критерии. Здесь применяется руководящий принцип, по которому формирующие ту или иную страту единицы должны быть в максимально возможной степени аналогичными в отношении переменных величин исследования для снижения вариантности в рамках каждой страты.

72. Вторым правилом стратификации является то, что каждая создаваемая страта должна в идеальном случае как можно больше отличаться от других. Следовательно, основным принципом формирования страт должна быть разнородность *между* стратами и однородность *внутри* страт. Таким образом, легко понять, почему городские и сельские районы зачастую формируются в качестве двух отдельных страт для обследования домашних хозяйств. Как указывалось выше, городское и сельское население отличается друг от друга по многим аспектам (вид занятости, источник и размер дохода, средний размер домохозяйства, уровень рождаемости и т. д.), в то время как лица, относящиеся к одной из этих подгрупп, обладают аналогичными характеристиками.

73. Однородность является полезным руководящим принципом для определения того, какое количество страт необходимо создать. Количество страт не должно превышать число поддающихся учету подгрупп населения в соответствии с определенным критерием, используемым для разграничения страт. Например, если какая-либо страна в административных целях разделена на восемь географических регионов, и при этом два из этих регионов весьма похожи друг на друга в отношении предмета предполагаемого обследования, надлежащий план выборки может быть достигнут путем создания семи страт (и объединяя два похожих региона). Никакого положительного эффекта не будет достигнуто в результате использования, например, 20 страт, если 10 из них могут предоставить одни и те же однородные подгруппы.

74. Важно отметить, что применительно к *пропорциональному* отбору результирующая выборка будет по крайней мере столь же точной, как простая случайная выборка такого же размера. Следовательно, стратификация дает повышение точности или надежности оценок обследования, при этом такое увеличение является максимальным в наиболее однородных стратах. Именно эта особенность стратифицированной выборки гарантирует, что даже неудачная стратификация[8] не нанесет ущерба оценкам обследования с точки зрения их надежности.

75. Другой важный аспект касается оценки ошибок выборки. Хотя выбор одной единицы из каждой страты достаточен для выполнения теоретических требований стратифицированной выборки, необходимо отбирать *как минимум две единицы* для того, чтобы результаты выборки можно было использовать при расчете ошибок выборки в оценках обследования.

76. Иногда возникает необходимость использования многих переменных величин в целях стратификации. В таких случаях мы должны руководствоваться следующими факторами: предпочтительно, чтобы используемые для стратификации переменные величины были не связаны друг с другом, но при этом связаны с переменной величиной обследования; нет необходимости добиваться полной завершенности и при формировании ячеек (более мелкие и наименее важные ячейки могут объе-

8 Неудачная стратификация имеет место при создании страт без необходимости, или когда некоторые элементы совокупности неправильно классифицированы и отнесены к неверной страте.

диняться); как правило, более заметного улучшения можно добиться путем использования более грубой разбивки многих переменных величин, чем более детальной разбивки одной переменной.

3.4.3. Неявная стратификация

77. Как указывалось выше, выбор имеющейся информации для создания страт зависит от целей измерения в ходе обследования. Для обследований домашних хозяйств, которые носят крупномасштабный и многоаспектный характер, особенно полезным методом является так называемая *неявная* стратификация. Тот факт, что ее важнейшим критерием являются географические особенности, как правило, служит достаточной основой для надлежащего распределения выборки по важным подгруппам населения, таким как городское и сельское население, административные регионы, этнические подгруппы, социально-экономические группы и т. д. В силу этого географического критерия неявная стратификация является весьма полезной даже тогда, когда предметом обследования служит единственная тема, например трудоспособное население, экономическая активность домашних хозяйств, измерение уровня бедности, здравоохранение или доходы и расходы. Этот метод настоятельно рекомендуется как по указанным выше причинам, так и в силу простоты его применения.

78. Для правильного применения неявная стратификация требует использования систематической выборки на первом этапе формирования выборки. Эта процедура проста в применении и предусматривает вначале расстановку файла ПЕВ в географической последовательности. Во многих странах такая последовательность, вероятнее всего, будет следующей: городское население в разбивке по провинциям и далее в пределах каждой провинции в разбивке по районам, затем сельское население в разбивке по провинциям и далее в пределах каждой провинции — по районам. Следующим шагом является систематический отбор ПЕВ из рассортированного файла. Систематический отбор производится либо путем равновероятностной выборки, либо, скорее всего, путем вероятностной выборки пропорционально размеру.

79. Как уже упоминалось, важным преимуществом неявной стратификации является то, что с ее помощью исключается необходимость в создании явных территориальных страт. Это, в свою очередь, устраняет необходимость распределять выборку между этими стратами, особенно при использовании пропорциональной выборки. Другим преимуществом является простота метода, рассмотренная в предыдущем пункте, поскольку этот метод требует всего лишь сортировки файла и применения надлежащего интервала(ов) выборки. Столь же легко можно применять непропорциональную выборку на первом уровне сортировки по географическому признаку. Например, если городское и сельское население представляет собой первый уровень, то достаточно понятной операцией выглядит применение различных долей выборки к городской и сельской частям населения. На рисунке 3.1 представлена схема неявной стратификации с систематической выборкой. Более подробно проблема формирования выборки с вероятностью пропорциональной размеру, рассматривается в разделе 3.6, ниже.

3.5. Гнездовая выборка

80. Изначально термин «гнездовая выборка» был введен для обозначения планов выборки, в которую входили все члены той или иной группы. Сами группы обозначались как кластеры или гнезда. Например, на первом этапе могла формироваться выборка школ, а на втором — классы. Если обследовались учащиеся каждого класса, тогда имелась гнездовая выборка классов. В обследованиях домашних хозяйств в качестве примера изначального понятия гнездовой выборки может служить отбор городских микрорайонов, в которых опросу в целях обследования подвер-

Рисунок 3.1
Распределение административных единиц в целях неявной стратификации

Городское население		
	Провинция 01	
		Район 01
		СУ 001
		СУ 002
		СУ 003
		СУ 004
		Район 02
		СУ 005
		СУ 006
		СУ 007
	Провинция 02	
		Район 01
		СУ 008
		СУ 009
		Район 02
		СУ 010
		СУ 011
		СУ 012
	Провинция 03 и т. д.	
Сельское население		
	Провинция 01	
		Район 01
		СУ 101
		СУ 102
		СУ 103
		СУ 104
		Район 02
		СУ 105
		СУ 106
		СУ 107
	Провинция 02	
		Район 01
		СУ 108
		СУ 109
		СУ 110
		СУ 111
		Район 02
		СУ 112
		СУ 113
		СУ 114
	Провинция 03 и т. д.	

гались все жители данного микрорайона. В последние годы, однако, термин «гнездовая выборка» широко используется для обозначения более общего плана обследований, имеющих предпоследний этап формирования выборки, для которой отбирались (и определялись) такие кластеры, как деревни, счетные участки переписи или городские микрорайоны. На последнем этапе формирования выборки проводится обследование не всех домохозяйств, а подгруппы выборки, состоящей из домохозяйств каждого отобранного кластера. Именно последний вариант применения данного термина, как правило, и используется в настоящем руководстве.

81. В обследованиях домашних хозяйств план выборки будет неизменно и по необходимости использовать ту или иную форму гнездовой выборки для сдерживания расходов на проведение обследования. Как указывалось выше, гораздо дешевле провести обследование, скажем, 1 000 домашних хозяйств по 50 адресам (20 домохозяйств на один кластер), чем 1 000 домохозяйств, произвольно отобранных из всего населения. К сожалению, формирование кластеров выборки снижает надежность выборки в связи с высокой вероятностью того, что лица, проживающие в одном кластере, склонны к однородности или к обладанию более или менее аналогичными характеристиками. Этот так называемый эффект гнездовой группировки необходимо компенсировать в плане выборки путем соразмерного увеличения размера выборки.

3.5.1. Характеристики гнездовой выборки

82. Гнездовая выборка значительно отличается от стратифицированной выборки по двум аспектам[9]. Что касается последней — в выборке представлены все страты, поскольку единицы отбираются в выборку из каждой страты. В гнездовой выборке производится отбор самих кластеров, следовательно, вошедшие в выборку кластеры представляют не вошедшие кластеры. Это первое характерное различие между стратифицированной и гнездовой выборками обусловливает и второе их различие. Как указывалось выше, в идеальном случае страты должны создаваться таким образом, чтобы быть внутренне однородными при внешней разнородности в плане подлежащих измерению параметров обследования. Для кластеров справедливым является прямо противоположный подход. С точки зрения точности выборки, лучше, чтобы кластеры были по возможности внутренне разнородными.

83. Тот факт, что в обследованиях домашних хозяйств кластерами практически всегда являются территориальные единицы, такие как деревни или части деревень, к сожалению, означает, что, как правило, невозможно добиться высокой степени разнородности в пределах одного кластера. И действительно, географически определяемые кластеры с большой вероятностью будут внутренне однородными, чем разнородными, в плане таких переменных показателей, как вид занятости (например, фермерская деятельность), уровень дохода и так далее. Следовательно, степень однородности кластеров применительно к тому или иному переменному показателю определяет, насколько «подверженным гнездовой выборке» может быть та или иная выборка. Чем в большей степени гнездовая группировка присутствует в выборке, тем ниже ее надежность.

3.5.2. Эффект схемы применительно к кластеру

84. Эффект гнездовой группировки в выборке частично измеряется с помощью эффекта схемы (*deff*). Тем не менее *deff* также отражает эффекты стратификации. Группа по составлению плана

9 Важно отметить, что стратификация и гнездовая выборка не являются конкурирующими друг с другом вариантами плана выборки, поскольку оба этих метода неизменно используются в выборке обследований домашних хозяйств.

выборки должна позаботиться о максимально возможном достижении в плане выборки оптимального баланса между минимизацией расходов и максимально возможной точностью данных. Это достигается, по мере возможности, путем сведения к минимуму или ограничения эффекта схемы. Для определения того, каким образом можно минимизировать или ограничить относящийся к гнездовой группировке элемент *deff,* полезно взглянуть на его математическое определение:

$$deff = 1 + \delta\,(\tilde{n} - 1), \qquad (3.10)$$

где δ — это внутриклассовая (или внутрикластерная) корреляция или, иными словами, та степень вероятности, с которой две единицы в кластере в сравнении с двумя единицами, произвольно отобранными из совокупности, будут иметь одно и то же значение; а $\tilde{n}$— это˜ число единиц обследуемой совокупности в данном кластере.

85. Строго говоря, уравнение (3.10) не является формулой расчета для *deff,* поскольку в нем не учитывается стратификация, а также еще один фактор, который вводится в случае неоднородности кластеров по размеру. Тем не менее, поскольку относящийся к гнездовой группировке элемент является доминирующим фактором в величине *deff,* он может применяться в качестве приблизительной формы расчета, показывая, в какой степени гнездовая группировка воздействует на план выборки и что можно сделать для ограничения такого воздействия.

86. Из приведенного выше математического выражения можно увидеть, что *deff* — это мультипликативная функция двух переменных: внутриклассовой корреляции, δ, и размера кластера, $\tilde{n}$. Следовательно, значение *deff* возрастает при увеличении как δ, так и $\tilde{n}$. Хотя составитель выборки не может контролировать величину внутриклассовой корреляции для любой искомой переменной, он/она все же может при формировании выборки скорректировать размер кластера в сторону его увеличения или уменьшения и, таким образом, в значительной мере контролировать эффект схемы.

Пример

Предположим, что некая совокупность имеет внутриклассовую корреляцию, равную 0,03, что является достаточно малой величиной для случая хронических заболеваний. Предположим также, что планирующие выборку специалисты обсуждают, следует ли использовать кластеры из 10 или 20 домохозяйств при общем размере выборки в 5000 домохозяйств. Предположим далее, лишь для упрощения примера, что все домохозяйства имеют одинаковый по составу размер — пять человек. Тогда величина $\tilde{\boldsymbol{n}}$ будет равна 50 для 10 домохозяйств и 100 для 20 домохозяйств. Простая замена в уравнении (3.4) дает примерную величину *deff* [1 + 0,03(49)] или 2,5 для плана кластеров из 10 домохозяйств, но 4,0 — для плана кластеров из 20 домохозяйств. Следовательно, эффект схемы будет примерно на 60 процентов выше для более крупного размера кластера. Затем группа по проведению обследования должна решить, какой из двух вариантов лучше — включить в выборку в два раза больше кластеров (500), используя вариант с 10 домохозяйствами для сохранения надежности на более приемлемом уровне, или выбрать более дешевый вариант из 250 домохозяйств, заплатив за это значительным увеличением вариантности выборки. Безусловно, можно рассматривать и другие варианты в промежутке между 10 и 20 домохозяйствами.

87. Существует несколько путей интерпретации эффекта схемы: как множитель, на который вариантность фактического плана выборки, используемого (предполагаемого к использованию) в ходе обследования, превышает вариантность простой случайной выборки (ПСВ) такого же размера; просто как измеритель того, насколько фактический план выборки хуже простой случайной выборки

с точки зрения точности; или показатель, отражающий, сколько дополнительных единиц выборки необходимо отобрать в намечаемом плане выборки по сравнению с простой случайной выборкой для достижения такого же уровня вариантности выборки. Например, *deff* величиной в 2,0 означает, что необходимо удвоить число единиц, подлежащих выборке, для достижения такой же надежности, какую имеет простая случайная выборка. В связи с этим крайне нежелательно иметь план выборки со значениями *deff*, значительно превышающими 2,5–3,0 для ключевых показателей.

3.5.3. Размер кластера

88. Ранее было отмечено, что составитель выборки не может контролировать корреляцию. Кроме того, в отношении большинства переменных показателей обследования практически не существует эмпирических исследований, в которых делалась бы попытка оценки величины таких корреляций. Внутриклассовая корреляция теоретически может варьироваться в пределах от –1 до +1, хотя сложно представить себе переменные показатели по домашних хозяйствам, которые имели бы отрицательное значение. В связи с этим единственная возможность, которой располагает составитель выборки для сдерживания величины *deff* на минимальном уровне, — это позаботиться о том, чтобы размеры кластеров были настолько малыми, насколько это позволяет бюджет. В таблице 3.3 указаны показатели *deff* для различных величин внутриклассовой корреляции и постоянного размера кластера.

89. Из таблицы 3.3 ясно видно, что размеры кластеров, превышающие 20, дают неприемлемые показатели *deff* (выше 3,0), если только при этом величина внутриклассовой корреляции не является весьма малой. При оценке цифр в таблице важно помнить, что $\tilde{n}$ обозначает число единиц в обследуемой совокупности, а не число домохозяйств. В этой связи величина $\tilde{n}$, которая должна использоваться, равняется числу домохозяйств в кластере, помноженному на среднее число лиц в обследуемой группе населения. Если обследуемой группой являются, например, женщины в возрасте 14–49 лет, то обычно на одно домохозяйство приходится примерно одна женщина из такой группы, и в этом случае размер кластера из *b* домохозяйств будет иметь приблизительно такое же число женщин в возрасте 14–49 лет. Иными словами, значения $\tilde{n}$ и *b* примерно равны для данной обследуемой группы, и таблица 3.3 применяется в данном неизменном виде. В следующем примере число домохозяйств и обследуемая совокупность в кластере не равны друг другу.

Таблица 3.3
Сравнение связанных с гнездовой группировкой элементов в эффекте схемы для различных величин внутриклассовой корреляции δ и размеров кластеров $\tilde{n}$

	δ						
$\tilde{n}$	0,02	0,05	0,10	0,15	0,20	0,35	0,50
5	1,08	1,20	1,40	1,60	1,80	2,40	3,00
10	1,18	1,45	1,90	2,35	2,80	4,15	5,50
20	1,38	1,95	2,90	3,85	4,80	6,65	10,50
30	1,58	2,45	3,90	5,35	6,80	11,15	15,50
50	1,98	3,45	5,90	8,35	10,80	18,15	25,50
75	2,48	4,70	8,40	12,10	15,80	26,90	38,00

Пример

Предположим, что изучаемой совокупностью является все население, как это было бы в случае обследования состояния здравоохранения для оценки числа острых и хронических заболеваний. Предположим далее, что в этом обследовании планируется применять кластеры из 10 домашних хозяйств. В этом случае величина $\tilde{n}$ будет в 10 раз больше размера среднего домохозяйства; если последний равен 5,0, то значение $\tilde{n}$ будет равно 50. Следовательно, 50 — это величина $\tilde{n}$, которая должна использоваться в таблице 3.3 для оценки возможного значения *deff*. Таблица 3.3 показывает, что значение *deff* будет весьма высоким, кроме случаев, когда величина δ примерно равна 0,02. Это означает, что гнездовая выборка, в которой запланировано использование кластеров всего лишь из 10 домохозяйств, даст весьма ненадежные результаты для такой характеристики, как инфекционные заболевания, поскольку последний показатель будет, вероятнее всего, иметь большую величину δ.

90. Этот пример демонстрирует, почему столь важно учитывать размер выборки при планировании обследования домашних хозяйств, особенно в отношении ключевых показателей, подлежащих измерению. Более того, следует иметь в виду, что размер кластера, заявленный в описании плана выборки, будет, как правило, означать число домохозяйств, в то время как размер кластера для целей оценки эффекта схемы должен вместо этого означать обследуемую группу(ы) населения.

3.5.4. Расчет эффекта схемы (*deff*)

91. Указанные аналитиками фактические величины *deff* для переменных показателей обследования могут быть рассчитаны уже после завершения обследования. Это требует оценки вариантности выборки для выбранных переменных (соответствующие методы рассматриваются в главе 7), а затем расчета для каждой переменной величины соотношения ее вариантности к вариантности простой случайной выборки из такого же общего размера выборки. Этот расчет дает оценку «полного» показателя *deff*, включая эффекты стратификации, а также вариантность размеров кластеров, а не только лишь элемента, связанного с гнездовой группировкой.

92. Квадратный корень из соотношения величин вариантности дает коэффициент среднеквадратической ошибки или так называемый показатель *deft*, который часто рассчитывается на практике и заносится в техническую документацию таких видов обследований, как Обследования в области народонаселения и здравоохранения (ОНЗ).

3.5.5. Число кластеров

93. Важно не упускать из виду, что размер кластера имеет большое значение и за пределами его воздействия на точность выборки и также играет важную роль применительно к общему размеру выборки, поскольку размер кластера определяет число различных адресов, которые необходимо посетить в ходе обследования. Такое существенное влияние на уровень затрат в ходе обследования — это, безусловно, именно та причина, по которой гнездовая выборка в первую очередь и применяется. Следовательно, выборка из 10 000 домашних хозяйств с кластерами размером в 10 домохозяйств каждый потребует 1 000 кластеров, в то время как кластеры из 20 домохозяйств потребуют только 500. Как подчеркивалось выше, чрезвычайно важно принимать во внимание факторы, как уровня затрат, так и точности при принятии решения по этой характеристике плана выборки.

3.6. Поэтапная выборка

94. В теории идеальный план выборки при обследовании домашних хозяйств предполагает формирование выборки в произвольном порядке из *n* домашних хозяйств среди надлежащим образом определенных страт, составляющих генеральную совокупность домохозяйств, *N*. Полученная таким способом стратифицированная случайная выборка даст максимальный уровень точности. Однако использование выборки такого вида является слишком дорогостоящим делом для целей практической реализации[10], как мы отметили выше при рассмотрении эффективности затрат, обеспечиваемой гнездовой выборкой.

3.6.1. Преимущества поэтапной выборки

95. Формирование выборки *в несколько этапов* имеет практические преимущества в плане самого процесса отбора. Это позволяет составителю выборки выделить с помощью последовательных шагов географические места проведения работ по обследованию, а именно по составлению списков и проведению опросов. Когда возникает необходимость в составлении списков по причине устаревшего инструментария выборки, для ограничения размеров района, по которому должны составляться списки, может вводиться этап отбора.

96. В случае гнездовой выборки, как правило, существуют не менее двух этапов процедуры отбора: первый этап — отбор кластеров и второй — отбор домашних хозяйств. В обследованиях домашних хозяйств кластеры всегда определяются как территориальные единицы того или иного вида. Если такие единицы достаточно малы, с точки зрения как территории, так и численности населения, и если имеется их обновленный, полный и точный список, из которого можно сформировать выборку, тогда для плана выборки будет достаточно и двух этапов. Если наименьшая из имеющихся территориальная единица слишком велика для эффективного использования в качестве кластера, будут необходимы три этапа отбора.

Пример

Предположим, что та или иная страна хочет в качестве кластеров определить счетные участки переписи (СУ), поскольку это — наименьшая существующая административно-территориальная единица. *Инструментарий* СУ (более подробно вопрос инструментария рассматривается в главе 4) уже разработан, поскольку вся страна разбита на *СУ*. Он является точным, поскольку каждое домохозяйство, по определению, находится в одном единственном *СУ*. Более того, такой инструментарий вполне пригоден для применения в том смысле, что он базируется на последней переписи населения, при условии что после этой переписи не было внесено никаких изменений в границы *СУ*. Предположим далее, что перепись проводилась два года назад. В связи с этим определяется необходимость в составлении более нового списка домашних хозяйств в *СУ*, вошедших в выборку, вместо того чтобы использовать полученный в результате переписи список домохозяйств двухлетней давности. Средний размер *СУ* составляет 200 домохозяйств, при том что желаемый размер кластера в целях проведения опросов определен в 15 домохозяйств на один кластер. Группа обследования проводит расчет, по которому стоимость составления списка из 200 домохозяйств ради тех 15, которые в конечном счете войдут в выборку (соотношение более 13 к 1), слишком велика. Затем составитель выборки принимает решение о проведении менее дорогостоящей опера-

10 Существует всего лишь несколько исключений, и они касаются стран с очень небольшой территорией, таких как Кувейт, где формирование случайной выборки домохозяйств будет связано с очень низкими транспортными расходами.

ции на местах, в рамках которой каждый вошедший в выборку *СУ* будет поделен на квадранты приблизительно равного размера примерно по 50 домохозяйств в каждом. Затем план выборки модифицируется в сторону отбора одного квадранта или *сегмента* из каждого вошедшего в выборку *СУ*, в котором и будет проводиться составление списка, в результате чего объем работы по составлению списка снизится на три четверти. При таком плане действий мы имеем три этапа: первый этап — отбор *СУ*, второй этап — отбор сегментов *СУ* и третий этап — отбор домашних хозяйств.

3.6.2. Использование фиктивных этапов

97. Часто при формировании выборки используются так называемые фиктивные этапы, с тем чтобы уйти от необходимости на предпоследнем этапе делать выборку из огромного списка единиц. Этот список может содержать такое количество единиц и может быть столь громоздким, что он станет практически неуправляемым в ходе трудоемких операций ручного отбора. Даже если этот список сделан в виде компьютерного файла, он может быть все же слишком большим для эффективного управления им при формировании выборки[11]. Фиктивные этапы позволяют сузить подсовокупности до более управляемых размеров, используя преимущества иерархической структуры административных единиц той или иной страны.

98. Например, для обследований сельского населения в Бангладеш деревни часто определяются в качестве единиц предпоследнего этапа формирования выборки. В Бангладеш существует более 100 000 деревень, и это, конечно, слишком много для управления процессом эффективного формирования выборки. Если план выборки предусматривает на предпоследнем этапе отбор, например, 600 деревень, то в случае Бангладеш в выборку будет включаться только одна деревня из 167. В целях сокращения размера файлов для формирования выборки может быть принято решение о формировании выборки в несколько этапов, используя иерархическую структуру единиц административно-территориального деления в Бангладеш, к которым относятся «таны» (thanas), «союзы» (unions) и деревни. Формирование выборки должно осуществляться поэтапно: сначала отбираются 600 танов при применении выборки с вероятностью, пропорциональной их размеру (этот метод подробно рассматривается в разделе 3.7). Затем из каждого вошедшего в выборку тана отбирается только один союз, снова при применении выборки с вероятностью, пропорциональной размеру, — таким образом в выборку войдут 600 союзов. На третьем этапе опять с применением выборки с вероятностью, пропорциональной размеру, одна деревня отбирается из каждого вошедшего в выборку союза, что в результате вновь даст 600 деревень. И наконец, выборка домохозяйств формируется из каждой отобранной деревни. Такая процедура в целом дает систематическую выборку всех домохозяйств по каждой отобранной деревне.

99. Описанная выше методология формирования выборки по сути является двухэтапной выборкой деревень и домохозяйств, хотя первоначально и были использованы два фиктивных этапа отбора танов и союзов, из которых затем были отобраны деревни. В этом случае необходимо математически пояснить фиктивный характер первых двух этапов путем изучения вероятностей на каждом этапе отбора, а также полной вероятности.

11 Тем не менее существует возможность сделать весьма большой компьютерный файл более приемлемым для формирования выборки, например, путем разбивки его на отдельные подфайлы для каждой страты или административной единицы (например, региона или провинции).

3.6.2.1. Первый этап отбора: таны

100. Таны отбираются на основе применения выборки с вероятностью, пропорциональной их размеру. На этом этапе вероятность определяется формулой:

$$P_1 = \left(am_t\right) \Big/ \sum m_t \ , \qquad (3.11)$$

где P_1 — это вероятность отбора данного тана; a — число подлежащих отбору танов (в нашем примере — 600), а m_t— число сельских домохозяйств[12] в t-ом тане в соответствии с использованным инструментарием выборки (например, последняя перепись населения).

101. Знаменатель $\sum m_t$ — это общее число сельских домохозяйств во всех танах страны. Необходимо отметить, что фактическое число отобранных танов может быть менее 600. Это может произойти в том случае, когда один или более танов отбираются дважды, так как эта возможность существует для любого тана, показатель размера которого превышает интервал выборки. Интервал выборки при отборе танов определяется уравнением $\sum m_t \div a$. Следовательно, если интервал выборки равен, скажем, 12 500, а данный тан содержит 13 800 домохозяйств, он будет автоматически отобран один раз и будет иметь вероятность 1 300/12 500 быть отобранным дважды (числитель дроби равен 13 800–12 500).

3.6.2.2. Второй этап отбора: союзы

102. На втором этапе один союз отбирается из каждого вошедшего в выборку тана вновь на основе применения *выборки с вероятностью, пропорциональной размеру*. На практике это достигается путем составления списка всех союзов в отобранном тане, суммируя показатели их размеров, m_u, и выбирая случайный номер между 1 и m_t, т. е. показателем размера вошедшего в выборку тана. Кумулянт, величиной которого является наименьший номер, равный или превышающий случайно выбранный номер, обозначает отобранный союз (либо для выявления отобранного союза может использоваться равнозначная методика). Если тот или иной тан на первом этапе был отобран дважды, то из него отбирается то же самое число союзов. Вероятность на втором этапе определяется формулой:

$$P_2 = (1)\left(m_u\right) \Big/ m_t \ , \qquad (3.12)$$

где P_2 — это вероятность отбора данного союза из вошедшего в выборку тана; (1) означает, что отбирается только один союз, а m_u — это число домохозяйств в u-ом союзе в соответствии с инструментарием.

12 Это — единица размера и она может в качестве варианта отображать численность населения таны при том условии, что используемое число совместимо со всеми единицами размера на каждом этапе.

3.6.2.3. Третий этап отбора: деревни

103. На третьем этапе одна деревня отбирается на основе применения *выборки с вероятностью, пропорциональной размеру*, из каждого вошедшего в выборку союза. Вероятность на третьем этапе определяется формулой:

$$P_3 = (1)\left(m_v\right)\Big/ m_u \ , \tag{3.13}$$

где P_3 — это вероятность отбора данной деревни из вошедшего в выборку союза; (1) означает, что отбирается только одна деревня, а m_v — это число домохозяйств в v-ой деревне в соответствии с инструментарием.

3.6.2.4. Четвертый этап отбора: домашние хозяйства

104. На четвертом этапе мы исходим из допущения о том, что инструментарий в виде списка домохозяйств имеется по каждой вошедшей в выборку деревне, а значит, на базе данных списков может быть сформирована систематическая выборка домохозяйств. Из каждой вошедшей в выборку деревни отбирается фиксированное число домашних хозяйств, и это число является заранее определенным размером кластера. Вероятность на четвертом этапе определяется формулой

$$P_4 = (b)\Big/ m_v \ , \tag{3.14}$$

где P_4 — это вероятность отбора данного домохозяйства из вошедшей в выборку деревни, а b — это фиксированное число домохозяйств, отбираемых в каждой деревне.

3.6.2.5. Полная вероятность отбора

105. Полная вероятность является произведением всех величин вероятности на каждом этапе и определяется формулой:

$$P = P_1\ P_2\ P_3\ P_4 \ . \tag{3.15}$$

Заменяя члены уравнения, получаем

$$P = \left[\left(am_t\right)\Big/\sum m_t\right]\left[(1)\left(m_u\right)/m_t\right]\left[(1)\left(m_v\right)/m_u\right]\left[b/m_v\right] = \left[(a)(b)\right]\Big/\sum m_t \ . \tag{3.16}$$

106. Следует отметить, что члены уравнения, P_2 и P_3, полностью взаимно исключаются, показывая фиктивный характер «четырехэтапной» процедуры отбора. Следовательно, хотя таны и союзы физически и «отбираются», они, тем не менее, служат исключительно для выявления местоположения вошедших в выборку деревень.

3.6.3. Двухэтапная схема

107. В последнее время много внимания уделяется использованию в развивающихся странах двухэтапной схемы планов выборки. Это — предпочтительный вариант плана выборки для Обследований по многим показателям с применением гнездовой выборки (ОМПГВ), проведенных Детским фондом Организации Объединенных Наций (ЮНИСЕФ) более чем в 100 странах, начиная с середины 1990-х годов. Кроме того, эти планы выборки преимущественно используются в Обследованиях в области народонаселения и здравоохранения (ОНЗ).

108. Как правило, двухэтапная схема состоит просто из выборки, сформированной с вероятностью, пропорциональной размеру, и состоящей из нескольких сотен территориальных единиц, которые были надлежащим образом стратифицированы на первом этапе. Может быть составлен обновленный список домохозяйств как единиц выборки первого этапа, в зависимости от наличия информации, касающейся адреса и/или местоположения таких домохозяйств, и в зависимости от степени обновленности такой информации. Затем, на втором этапе следует систематическая выборка фиксированного числа домохозяйств. Территориальными единицами, обычно именуемыми «кластерами», как правило, являются деревни или счетные участки переписи в сельской местности и микрорайоны в городах.

109. Описанная выше двухэтапная схема привлекательна с многих точек зрения, но в основном в силу ее простоты. *В целях снижения возможности ошибок регистрации данных при проведении выборочного обследования всегда предпочтительно при составлении плана выборки стремиться к упрощению, нежели к усложнению.* Двухэтапная схема имеет целый ряд полезных характеристик, которые делают ее сравнительно простой и предпочтительной к использованию. Например:

- Как указывалось ранее, такой план выборки является самовзвешенным (все домохозяйства отбираются в выборку с равной вероятностью) или приблизительно самовзвешенным (вопрос различий между вероятностной выборкой, пропорциональной размеру, и вероятностной выборкой, пропорциональной предполагаемому размеру, рассматривается в разделах 3.7.1 и 3.7.2).
- Кластеры, определяемые в виде *СУ* или городских микрорайонов, в большинстве стран имеют удобный размер (не слишком большой), особенно в случае необходимости составления обновленного списка домохозяйств до окончательного этапа отбора.
- *СУ*, городские микрорайоны и большинство деревень обычно наносятся на карту либо для проведения работ переписи, либо в иных целях и имеют четкую делимитацию границ.

3.7. Выборка с вероятностью, пропорциональной размеру, и с вероятностью, пропорциональной предполагаемому размеру

110. В разделе 3.5 приведен пример, в котором при отборе кластеров для выборки применялась прежде всего выборка с вероятностью, пропорциональной размеру. В настоящем разделе метод *выборки с вероятностью, пропорциональной размеру,* рассматривается более подробно.

3.7.1. Выборка с вероятностью, пропорциональной размеру

111. Применение метода *выборки с вероятностью, пропорциональной размеру*, позволяет составителю выборки осуществлять более жесткий контроль над конечным размером выборки в обследованиях методом гнездовой выборки. В случаях, когда все кластеры имеют одинаковый

или примерно одинаковый размер, не существует никаких преимуществ в применении *выборки с вероятностью, пропорциональной размеру*. Предположим, например, что каждый микрорайон данного города содержит точно 100 домохозяйств, и при этом требуется выборка из 1 000 домохозяйств, распределенных по 50 вошедшим в выборку городским микрорайонам. Наиболее очевидным планом выборки будет формирование *ПСВ* или, иными словами, равновероятностной выборки из 50 микрорайонов, а затем систематический выбор лишь одного из пяти домохозяйств из каждого микрорайона (также методом равновероятностной выборки). Результатом будет выборка ровно 20 домохозяйств на каждый микрорайон, т. е. всего 1 000 домохозяйств. В этом случае уравнение отбора будет следующим:

$$p = (50/M)(1/5)\ ,$$

где p — это вероятность отбора домохозяйства; $(50/M)$ — вероятность отбора микрорайона; M — общее число микрорайонов в городе, а (1/5) — вероятность отбора того или иного домохозяйства из определенного вошедшего в выборку микрорайона.

112. Значение P сокращается до $10/M$. Поскольку M является постоянной величиной, общая вероятность отбора для каждого вошедшего в выборку домохозяйства равна значению 10, деленному на число микрорайонов, M.

113. Тем не менее в реальных ситуациях микрорайоны или иные территориальные единицы, которые могут использоваться в качестве кластеров для обследований домашних хозяйств, редко имеют столь постоянные размеры. Применительно к приведенному выше примеру, они могут варьироваться по размеру в пределах, допустим, от 25 до 200. Равновероятностная выборка микрорайонов может привести к неудачной выборке из преимущественно малых или, напротив, больших микрорайонов. Результатом в этом случае может стать общий размер выборки, существенно отличающийся от заданных 1 000 домохозяйств, которые приведены в этом примере. Одним из методов снижения возможности появления варьирующихся в широких пределах размеров выборки является создание страт, исходя из размера кластеров, и формирование выборки из каждой страты. Этот метод, как правило, не рекомендуется, поскольку он может снизить или усложнить использование других аспектов стратификации при формировании выборки. Предпочтительным решением в этом случае является применение выборки с вероятностью, пропорциональной размеру, поскольку она позволяет более жестко контролировать конечный размер выборки без необходимости введения стратификации по размеру.

114. Для иллюстрации применения выборки с вероятностью, пропорциональной размеру, мы начнем с выбора приведенного выше уравнения, однако, выраженного в более формализованном виде для двухэтапной схемы[13] следующим образом:

$$P(\alpha\beta) = P(\alpha)P(\beta|\alpha), \tag{3.17}$$

где $P(\alpha\beta)$ — это вероятность отбора домохозяйства β в кластере α; $P(\alpha)$ — вероятность отбора кластера α; $P(\beta|\alpha)$ — условная вероятность отбора домохозяйства β на втором этапе при условии, что кластер α был отобран на первом этапе.

13 См. (Kalton, 1983, pp. 38-47), где подробно изучается данное понятие и рассматривается метод выборки с вероятностью, пропорциональной размеру.

115. Для фиксирования общего размера выборки с точки зрения числа домохозяйств нам необходима *равновероятностная* выборка *n* домохозяйств из совокупности в *N* домохозяйств. Следовательно, общая доля выборки составит *n/N*, что равняется $P(\alpha\beta)$, как определено ниже. Далее, если число подлежащих включению в выборку кластеров обозначено как *a*, тогда в идеальном случае нам необходимо отобрать *b* домохозяйств из каждого кластера вне зависимости от размеров отобранных кластеров. Если мы определяем m_i как размер *i*-го кластера, тогда нам нужно, чтобы значение $P(\beta|\alpha)$ было равно b/m_i. Следовательно,

$$P(\alpha\beta) = [P(\alpha)]\left[b/m_i\right].$$

Поскольку $n = ab$, мы имеем

$$ab/N = [P(\alpha)]\left[b/m_i\right].$$

Решив последнее уравнение в отношении $P(\alpha)$, мы получаем

$$P(\alpha) = (a)(m_i)/N. \qquad (3.18)$$

116. Следует отметить, что $N = \sum m_i$, и, таким образом, вероятность отбора кластера пропорциональна его размеру. Уравнение формирования выборки из единиц первого этапа с *вероятностью, пропорциональной размеру,* в которой конечные единицы, тем не менее, отбираются с равной вероятностью, будет иметь следующий вид:

$$P(\alpha\beta) = \left[(a)(m_i)/\sum m_i\right]\left[b/m_i\right] \qquad (3.19)$$

$$= \left[(ab)/\sum m_i\right] \qquad (3.20)$$

117. Получаемый таким методом план выборки является самовзвешенным, что можно увидеть из уравнения (3.19), поскольку все члены уравнения являются постоянными величинами; напомним, что, хотя m_i является переменной величиной, сумма $\sum m_i$ — это константа, равная *N*. Приводимый ниже рисунок 3.2 может служить примером того, как сформировать выборку кластеров с *вероятностью, пропорциональной размеру.*

118. Что касается физического формирования выборки, следует отметить, что на рисунке 3.2 интервал выборки, *I*, последовательно добавляется к произвольно выбранной точке отсчета (R*S*) семь раз (или *a*–1 раз, где *a* — это число подлежащих отбору кластеров). Получаемые в результате числа отбираемых единиц составляют 311,2 (что является значением *RS*), 878,8; 1 446,4; 2 014,0; 2 581,6; 3 149,2; 3 716,8 и 4 284,4. Кластером, включаемым в выборку в соответствии с этими восемью числами отбираемых единиц, в каждом случае является тот кластер, чей суммарный показатель размера представляет собой наименьшую величину, равную или превышающую число отбираемых единиц. Таким образом, кластер 03 выбран в силу того, что 377 — это наименьший кумулянт, равный или превышающий величину 311,2, а кластер 26 выбран в силу того, что 3 744 — это наименьший кумулянт, равный или превышающий величину 3 716,8.

119. Несмотря на то что данная иллюстрация не очень убедительно демонстрирует эту закономерность (поскольку отобрано только восемь кластеров), выборка с вероятностью, пропорциональной размеру, имеет тенденцию отбирать скорее более крупные, нежели более мелкие кластеры. Это

Рисунок 3.2
Пример систематической выборки кластеров с вероятностью, пропорциональной размеру

Кластер/число ПЕВ	Показатель размера (*число домохозяйств*)	Совокупный показатель	Отбор для формирования выборки
001	215	215	
002	73	288	
003	89	377	311,2
004	231	608	
005	120	728	
006	58	786	
007	99	885	878,8
008	165	1 050	
009	195	1 245	
010	202	1 447	1 446,4
011	77	1 524	
012	59	1 583	
013	245	1 828	
014	171	1 999	
015	99	2 098	2 014,0
016	88	2 186	
017	124	2 310	
018	78	2 388	
019	89	2 477	
020	60	2 537	
021	222	2 759	2 581,6
022	137	2 896	
023	199	3 095	
024	210	3 305	3149,2
025	165	3 470	
026	274	3 744	3 716,8
027	209	3 953	
028	230	4 183	
029	67	4 250	
030	72	4 322	4 284,4
031	108	4 430	
032	111	4 541	

Инструкции по выборке: отберите 8 ПЕВ (кластеров) из 32 в данной совокупности, используя вероятность, пропорциональную размеру; в этом случае интервал выборки (*I*) будет равен 4 541/8, или 5 67,6, где 4 541 — это общий совокупный размер всех кластеров, а 8 — это число подлежащих отбору кластеров; произвольно выбранная точка отсчета (RS) — это случайное число между 0,1 и 567,6, выбранное по таблице случайных чисел; в данной иллюстрации RS = 311,2.

может быть очевидно, поскольку из формулы (3.17) следует, что вероятность отбора кластера пропорциональна его размеру; таким образом, кластер, содержащий 200 домохозяйств, имеет в два раза более высокую вероятность быть отобранным, нежели кластер из 100 домохозяйств. В связи с этим следует отметить, что один и тот же кластер может быть отобран более одного раза в случае, если его размер превышает интервал выборки, *I*. Однако ни один из кластеров на рисунке не соответствует этому условию, однако если это произойдет, то число домохозяйств, подлежащих отбору из такого кластера, должно удвоиться при двух «попаданиях», утроиться при трех попаданиях и т. д.

3.7.2. Выборка с вероятностью, пропорциональной предполагаемому размеру

120. Описанная в предыдущем разделе выборка *с вероятностью, пропорциональной размеру*, является в некоторой степени идеальным вариантом и в большинстве случаев не может быть осуществлена на практике. Это обусловлено тем, что показатель размера, используемый для установки уровня вероятности отбора кластера на первом этапе, зачастую не является *фактическим* показателем размера при формировании выборки домохозяйств на втором этапе.

121. В обследованиях домашних хозяйств в качестве показателя размера, обычно принимаемого на первом этапе отбора первичных единиц выборки или кластеров, выступает численность домохозяйств (или населения) из последней переписи населения. Даже если перепись проводилась совсем недавно, фактическая численность домохозяйств на момент переписи скорее всего будет отличаться, пусть и ненамного. Исключением, однако, является ситуация, когда отбор домохозяйств на втором этапе ведется напрямую из того же инструментария, который был использован для определения показателей размера (более подробно инструментарий выборки рассматривается в главе 4).

Пример

Предположим, что обследование домашних хозяйств проводится через три месяца после завершения переписи населения. Вместо составления нового списка домохозяйств в отобранных кластерах группа по проведению обследования принимает решение использовать на втором этапе двухэтапной выборки список переписи по домохозяйствам, поскольку достоверно считает, что такой список переписи является обновленным и точным для любых практических целей. На первом этапе формируется выборка деревень с использованием полученных в результате переписи данных о численности домохозяйств в качестве показателя размера для каждой деревни. Для каждой вошедшей в выборку деревни показатель размера, m_i, идентичен фактическому числу домохозяйств, из которых должна формироваться выборка. Следовательно, если будет выбрана деревня А и если в ней по данным переписи находится 235 домохозяйств, то список, из которого будет формироваться выборка домохозяйств для обследования, также будет содержать 235 домохозяйств.

122. Однако многие обследования домашних хозяйств, базирующиеся на инструментариях переписи, проводятся спустя много месяцев, а иногда и лет после переписи (вопрос обновления инструментариев выборки подробно рассматривается в главе 4). В таких обстоятельствах зачастую принимается решение о проведении соответствующей работы на местах по подготовке нового списка домохозяйств в тех кластерах, которые отобраны для включения в выборку на ее первом этапе. Затем из такого обновленного списка формируется выборка домохозяйств для проведения обследования.

123. В приведенном выше примере данные переписи по численности домохозяйств выступают в качестве показателя размера, m_i, который использовался для отбора кластера. Однако фактический список, из которого формируется выборка домохозяйств, будет иным. Он будет безусловно

иметь другой показатель размера, который в определенной степени будет зависеть от величины промежутка времени между датами проведения переписи и подготовки списка обследования. Отличия возникнут в результате миграции населения в данный кластер и из него, строительства новых и сноса старых зданий, возникновения новых отдельных домохозяйств в случае брака (иногда в одной и той же жилищной единице с родителями) или случаев смерти. Когда выборка формируется *с вероятностью, пропорциональной предполагаемому размеру*, ее вероятность из уравнения отбора будет следующей:

$$P(\alpha\beta) = \left[(a)(m_i)/\sum m_i\right]\left[b/m_i'\right], \qquad (3.21)$$

где m_i' — это численность домохозяйств в соответствии с составленным списком, а другие члены уравнения определяются, как указывалось ранее.

124. Поскольку m_i' и m_i будут скорее всего отличаться в большинстве, если не во всех кластерах выборки, при расчете вероятности отбора (a, следовательно, и веса или, иными словами, обратной функции вероятности) необходимо учитывать эти отличия. Как показывает уравнение 3.20, каждый кластер будет иметь различный вес, не позволяя, таким образом, создать самовзвешенный план выборки.

125. Путем использования точных весовых коэффициентов для компенсации различий между показателями размера по данным переписи и обследования можно добиться неискаженных результирующих оценок обследования. Отсутствие надлежащей коррекции весовых коэффициентов дает смещенные оценки, причем величина погрешностей будет без сомнения возрастать по мере увеличения интервала времени между датами переписи населения и обследования. Необходимо отметить, однако, что при наличии между m_i' и m_i лишь небольших различий, выборка становится практически самовзвешенной, и при определенных обстоятельствах[14] целесообразно получать оценки обследования без взвешивания, поскольку погрешности будут весьма незначительными. Однако до принятия решении о данной схеме действий очень важно определить величины m_i' и m_i по каждому кластеру, чтобы эмпирически оценить, действительно ли различия минимальны.

126. Существует альтернативная стратегия, которую можно использовать для отбора домохозяйств на последнем этапе в тех случаях, когда применяется выборка с вероятностью, пропорциональной размеру, т. е. выборка, которая фактически является самовзвешенной. Эта стратегия предусматривает отбор домохозяйств с переменной долей выборки в рамках каждого кластера в зависимости от его фактического размера (эта стратегия рассматривается в следующем разделе).

3.8. Варианты формирования выборки

127. В данном разделе рассматриваются некоторые из многочисленных вариантов, которые можно применять при составлении надлежащей выборки для обследований общего назначения в отношении домашних хозяйств, при этом основное внимание уделяется стратегиям отбора на предпоследнем и последнем этапах, поскольку именно для этих этапов имеется несколько вариантов стратегий. Кроме того, рассматривается вопрос выбора между *равновероятностной выборкой*

14 Такая стратегия применима в тех обследованиях, в которых оценки ограничены долями, соотношениями и коэффициентами; однако в тех обследованиях, в которых необходима оценка суммарных или абсолютных величин, необходимо использование весовых коэффициентов вне зависимости от того, является ли или нет выборка самовзвешенной, приближенно самовзвешенной или несамовзвешенной.

или *выборкой с вероятностью, пропорциональной размеру,* для кластеров на предпоследнем этапе, а также между выборками домохозяйств с равномерной долей или с постоянным размером на последнем этапе. Далее в разделе дается резюме предыдущих разделов по таким вопросам, как контроль размера выборки, сравнение преимуществ и недостатков самовзвешенного и несамовзвешенного плана выборки, а также по другим вопросам типа рабочей нагрузки регистраторов. Кроме того, в разделе рассматриваются широко применяемые в настоящее время конкретные планы выборки, такие как использовались для Обследования в области народонаселения и здравоохранения и проведенного ЮНИСЕФ в середине десятилетия Обследования по многим показателям с применением гнездовой выборки. Эти схемы дают дополнительные варианты, достойные изучения, включая применение компактных (полных) и некомпактных кластеров.

3.8.1. Равновероятностная выборка, выборка с вероятностью, пропорциональной размеру, выборка с постоянным размером и с фиксированной долей

128. В таблице 3.4 возможные для применения схемы создают основу для обсуждения процедур, условий, преимуществ и недостатков различных планов выборки.

Таблица 3.4
Альтернативные планы выборки: два последних этапа отбора

Отбор единиц предпоследнего уровня	Кластер постоянного размера (число домохозяйств)	Фиксированная доля отбора в каждом кластере
Выборка с вероятностью, пропорциональной размеру	План 1	План 2 [не рекомендуется]
Выборка с вероятностью, пропорциональной предполагаемому размеру	План 3	План 4 [не рекомендуется]
Равновероятностная выборка	План 5	План 6

129. Мы уже рассмотрели, что выборка с вероятностью, пропорциональной размеру, первичных единиц выборки или кластеров является более точным способом контроля конечного размера выборки, чем *равновероятностная* выборка, и это является ее главным преимуществом, особенно в том случае, если кластеры в широких пределах варьируются по числу входящих в каждый из них домохозяйств. Контроль над размером выборки важен не только в плане его влияния на расходы, но также и в плане того, что такой контроль позволяет руководителю обследования точно спланировать рабочую нагрузку регистраторов еще до начала проведения операций в рамках обследования. С другой стороны, *равновероятностная* выборка проще в осуществлении, нежели выборка *с вероятностью, пропорциональной размеру*, и ее применение имеет смысл в случаях, когда показатели размера кластеров примерно равны или отличаются незначительно. В практическом плане в случаях, когда фактический показатель размера отличается от указанного в инструментарии, вместо выборки с вероятностью, пропорциональной размеру, следует применять выборку с вероятностью, пропорциональной предполагаемому размеру.

130. Отбор фиксированного числа домохозяйств в каждом кластере выборки имеет два весьма важных преимущества: во-первых, точно контролируется размер выборки; во-вторых, этот метод предоставляет руководителю обследования возможность точно распределить рабочую нагрузку между регистраторами и, при их желании, выровнять эту нагрузку. При этом, однако, выборка с постоянным размером связана с определенными сложностями, поскольку она требует расчета различных интервалов выборки для каждого кластера. Применение различных интервалов выборки может вызвать путаницу и послужить источником ошибок. Тем не менее в этой схеме

присутствует «встроенный» механизм контроля качества, поскольку число подлежащих отбору домохозяйств известно заранее. И все же указанные сложности могут привести к снижению эффективности из-за потерь времени на исправление ошибок отбора.

131. Формирование выборки с постоянным размером по определению требует составления списка домашних хозяйств, на основе которого отобранные домохозяйства могут обозначаться и выявляться. Наиболее часто в качестве такого списка выступает текущий список, подготовленный в рамках проводимых на местах подготовительных операций для обследования. Целесообразно организовать в центральном офисе проведение отбора подлежащих включению домохозяйств, причем это должно быть сделано другим сотрудником, а не самим составителем списка, для сведения к минимуму субъективности процедуры отбора.

132. В качестве альтернативы домашние хозяйства могут отбираться, исходя из фиксированной доли в каждом кластере. В этом случае отбор будет проще и менее подвержен ошибкам. Преимуществом при работе на местах является то, что такая выборка может осуществляться в процессе обхода регистратором кластера в целях составления обновленного списка домохозяйств. Это достигается путем разработки формуляра списка, включающего заранее подготовленные графы для выявления отобранных для выборки домохозяйств. Тот факт, что составление списка и формирование выборки могут осуществляться в ходе одного посещения, дает очевидную экономию затрат; однако такой подход имеет ряд существенных недостатков.

133. Одним из недостатков выборки с фиксированной долей является слабый контроль над размером выборки или рабочей нагрузкой регистраторов, за исключением случаев, когда *показатели размера* каждого кластера примерно равны. Другим, более серьезным недостатком является часто имеющий место субъективный выбор в результате того, что регистраторам доверяется фактический отбор домашних хозяйств для выборки или, иными словами, выявление тех домохозяйств, которые должны включаться в графы формуляров в списке выборки. Были проведены многочисленные исследования, которые продемонстрировали, что при контроле над ситуацией регистраторы имеют тенденцию отбирать более мелкие по размеру домохозяйства, и это подводит к выводу о том, что регистраторы для снижения своей рабочей нагрузки могут преднамеренно или неосознанно выбирать домохозяйства с меньшим числом респондентов.

134. Такой план является самовзвешенным и не зависит от конкретного набора процедур выборки на каждом этапе. Следовательно, двухэтапный план, объединяющий выборку кластеров *с вероятностью, пропорциональной размеру*, и выборку домохозяйств с постоянным размером, является самовзвешенным, в то время как комбинация выборки *с вероятностью, пропорциональной размеру*, и выборки с фиксированной долей самовзвешенной не является. Ниже рассматривается вопрос о том, какие из планов выборки в таблице 3.4 являются самовзвешенными.

3.8.1.1. План 1: Вероятность, пропорциональная размеру, постоянный размер кластера

Условия:

- различные показатели размера для совокупности кластеров;
- домохозяйства отбираются из тех же списков (например, список переписи домохозяйств), которые используются для *определения показателей размера.*

Преимущества:

- контроль над общим размером выборки и, следовательно, затратами;
- контроль над рабочей нагрузкой регистраторов;
- самовзвешенная выборка.

Недостатки:

- *выборка с вероятностью, пропорциональной размеру*, несколько сложнее в применении, чем *равновероятностная выборка;*
- различные доли отбора для домохозяйств по каждому кластеру, что может стать источником ошибок.

3.8.1.2. План 2: Вероятность, пропорциональная размеру, выборка с фиксированной долей

135. Не существует приемлемых условий, при которых возможно использование такого плана выборки. Если кластеры различаются по размеру, тогда надлежащим планом выборки будет использование выборки с вероятностью, пропорциональной размеру, совместно с кластерами постоянного размера. Если кластеры имеют приблизительно равные размеры, тогда подходящей схемой является выборка с фиксированной долей, но при этом сами кластеры должны отбираться с помощью равновероятностной выборки.

3.8.1.3. План 3: Вероятность, пропорциональная предполагаемому размеру, постоянный размер кластера

Условия:

- различные *показатели размера* для совокупности кластеров;
- домохозяйства отбираются из новых списков, обновляющих списки инструментария для установления первоначального *показателя размера.*

Преимущества:

- контроль над общим размером выборки и, следовательно, затратами;
- контроль над рабочей нагрузкой регистраторов;
- является более точной, чем выборка с вероятностью, пропорциональной размеру, на основе данного инструментария, поскольку обновлены списки домохозяйств.

Недостатки:

- выборка с вероятностью, пропорциональной размеру, несколько сложнее в применении, чем метод равновероятностной выборки;
- различные доли отбора для домохозяйств по каждому кластеру, что может стать источником ошибок;
- несамовзвешенная выборка.

3.8.1.4. План 4: Равновероятностная выборка, с фиксированной долей

136. Не существует приемлемых условий, при которых возможно использование плана 4 по причинам, изложенным для плана 2.

3.8.1.5. План 5: Равновероятностная выборка, постоянный размер кластера

Условия:

- *показатели размера* для совокупности кластеров примерно равны или варьируются в минимальных пределах.

Преимущества:

- контроль над общим размером выборки (но несколько слабее, чем в плане 1) и, следовательно, затратами;
- контроль над рабочей нагрузкой регистраторов, но также несколько слабее, чем в плане 1;

- равновероятностную выборку легче применять, нежели выборку с вероятностью, пропорциональной размеру, или выборку с вероятностью, пропорциональной предполагаемому размеру.

Недостатки:

- различные доли отбора для домохозяйств по каждому кластеру, что может стать источником ошибок;
- несамовзвешенная выборка.

3.8.1.6. План 6: Равновероятностная выборка, с фиксированной долей

Условия:

- *показатели размера* в совокупности кластеров практически равны.

Преимущества:

- самовзвешенная выборка;
- предельная простота формирования выборки на обоих этапах.

Недостатки:

- слабый контроль над общим размером выборки с негативными последствиями для затрат и надежности данных, особенно в случае, если текущий показатель размера существенно отличается от показателя размера инструментария; а также негативные последствия для надежности данных, если выборка получается значительно меньше, чем планировалось;
- слабый контроль над рабочей нагрузкой регистраторов.

3.8.2. Обследование в области народонаселения и здравоохранения (ОНЗ)

137. Несмотря на то что в Обследовании в области народонаселения и здравоохранения (ОНЗ) основной упор делается на женщинах детородного возраста, его план выборки подходит для обследований общего назначения.

138. В пособии по вопросам выборки Обследования в области народонаселения и здравоохранения, которое, начиная с 1984 года, широко используется во многих развивающихся странах, продвигается идея применения *плана стандартных сегментов*[15] из-за его удобства и практичности. Стандартный сегмент определяется своим размером и составляет, как правило, 500 человек. Каждой территориальной единице страны, входящей в инструментарий выборки, присваивается показатель размера, рассчитываемый как численность населения этой единицы, деленная на 500 (или любой иной размер стандартного сегмента, принятый в данной стране). Результирующая величина, округленная до ближайшего целого числа, является числом стандартных сегментов в данной территориальной единице.

139. Выборка территориальных единиц с вероятностью, пропорциональной размеру, формируется с использованием стандартных сегментов в качестве показателя размера. Поскольку теми территориальными единицами, которые обычно используются на данном этапе выборки, выступают счетные участки (СУ), городские микрорайоны или деревни, показатель размера для большинства из них равняется единице или двум. В любой отобранной территориальной единице с *показателем размера*, превышающим единицу, проводится картографирование, в ходе которого создаются территориальные сегменты, при этом число таких сегментов равняется *показателю*

15 План стандартных сегментов также использовался в программе обследований 1980–1990-х годов Панарабского проекта по проблемам развития детей (ПАПЧАЙЛД) — см. Лига арабских государств (1990 год).

размера. Таким образом, территориальная единица выборки с *показателем размера* 3 будет разделяться на карте на три сегмента примерно равного размера в тех пределах, в которых позволяют ее естественные границы, с точки зрения численности населения в каждом сегменте (в отличие от размера ее территории).

140. Каждая территориальная единица с *показателем размера*, равным 1, автоматически включается в выборку, а во всех остальных единицах произвольно отбирается один сегмент с помощью равновероятностной выборки. Затем проводится обход всех сегментов выборки, включая отобранные автоматически, для получения обновленных списков домохозяйств. Постоянная часть (доля) домохозяйств систематически отбирается из каждого вошедшего в выборку «кластера» для последующего проведения опроса в рамках ОНЗ. Поскольку все сегменты имеют примерно равный размер, процедура выборки представляет собой двухэтапный план равновероятностной выборки сегментов и домохозяйств.

141. Используемый в ОНЗ план стандартных сегментов близок к указанному выше плану 6: равновероятностная выборка кластеров и выборка домохозяйств с фиксированной долей в рамках вошедших в выборку кластеров (также равновероятностную). Тем не менее благодаря процедуре разбивки на стандартные сегменты план ОНЗ не имеет тех серьезных недостатков, которые присущи плану 6: как общий размер выборки, так и рабочая нагрузка регистраторов, контролируется практически со 100-процентной точностью.

142. Важным преимуществом плана стандартных сегментов является значительное сокращение объема работ по составлению списков на предпоследнем этапе отбора. Для каждой территориальной единицы, состоящей из s сегментов, объем работ по составлению списков сокращается на $1/s$ (при наличии только одного сегмента объем работ не сокращается). Например, если данная территориальная единица содержит четыре сегмента, объем работ по составлению списков будет равен только одной четвертой части того объема работ, который потребуется для всей данной территориальной единицы. В связи с этой особенностью сократятся и затраты на подготовку выборки.

143. Даже при условии снижения затрат на составление списков, сокращение объема работ несет в себе определенные издержки. Недостатком плана стандартных сегментов является то, что для сегментов с показателем размера более 1 должно быть проведено картографирование. Такая работа по составлению карт может быть весьма трудоемкой и дорогостоящей, она требует особой подготовки и подвержена ошибкам. Тот факт, что естественные границы далеко не всегда четко показаны, препятствует обоснованной делимитации границ сегментов в рамках территориальной единицы. Этот недостаток усложняет работу регистраторов, которые позднее посещают такой сегмент, по установлению точного местоположения того или иного отобранного домохозяйства. Тем не менее эту проблему можно несколько упростить путем включения в список фамилии главы домохозяйства еще на этапе составления такого списка, и в этом случае плохо обозначенные границы приносят меньше неудобств.

3.8.3. Модифицированный план с применением гнездовой выборки: обследование по многим показателям с применением гнездовой выборки (ОМПГВ)

144. Специалисты-практики по проведению обследований зачастую жалуются на большие затраты и потери времени на составление списка домохозяйств в кластерах, включенных в выборку на предпоследнем этапе. Составление списков, как правило, требуется для большинства обследований — включая, как упоминалось выше, метод с использованием плана стандартных сегментов ОНЗ — для получения обновленного списка домохозяйств, из которого они должны отбираться для проведения опросов в рамках обследования. Это особенно важно в случаях, когда инстру-

ментарий выборки составлен более года назад. *На работы по составлению списков приходится значительная доля расходов и процедур обследования, что зачастую упускается из виду на этапах как планирования бюджета, так и составления графика обследования*. Составление списков требует посещения районов обследования в дополнение к посещениям, необходимым для проведения опроса. Более того, число включаемых в список домохозяйств нередко в 5–10 раз превышает число домохозяйств, отобранных для опроса. Предположим, например, что план выборки предусматривает отбор 300 ПЕВ с размером кластера в 25 домохозяйств для проведения опроса в общей сложности 7 500 домохозяйств. Если средняя *ПЕВ* предпоследнего уровня содержит 150 домохозяйств, тогда необходимо составить список из 45 000 домохозяйств.

145. Стратегия выборки, использованная для Обследования с применением гнездовой выборки в рамках Расширенной программы вакцинации (РПВ) (Всемирная организация здравоохранения, 1991 год), была разработана Центрами контроля заболеваний и частично Всемирной организацией здравоохранения (ВОЗ) для того, чтобы избежать расходов и потерь времени, связанных с составлением списков. Обследование с применением гнездовой выборки РПВ, предназначенное для оценки степени охвата детей вакцинацией, широко использовалось многими развивающимися странами в течение более двадцати лет. Важная статистическая проблема (Turner, Magnani and Shuaib, 1996) касается методологии формирования выборки. Методология обследований с помощью гнездовой выборки использует квотную выборку на втором этапе отбора, даже несмотря на то, что единицы первого этапа (деревни или микрорайоны) обычно отбираются в соответствии с принципами вероятностной выборки. Часто используемый, хотя и с вариациями, метод квотной выборки предусматривает начало проведения опросов в рамках обследования в некоторой центральной точке в отобранной деревне, а затем — продвижение в случайно выбранном направлении, продолжая опрос домохозяйств до выполнения установленной квоты. Согласно варианту Обследования с применением гнездовой выборки РПВ, посещение домохозяйств продолжается до тех пор, пока не будут найдены семеро детей из целевой возрастной группы. Хотя преднамеренное искажение и отсутствует при использовании таких методов, в течение длительного времени по ним высказывались различные критические замечания со стороны многих специалистов-статистиков, в том числе в работах Калтона (1987 год), Скотта (1993 год) и Беннетта (1993 год). Основной упор в критике делается на то, что эта методология не дает вероятностную выборку (см. раздел 3.2 по вопросу сравнения между вероятностной выборкой и другими методами выборки для обсуждения причин того, почему вероятностная выборка является рекомендуемым подходом для обследований домашних хозяйств).

146. Один из вариантов метода Обследования с применением гнездовой выборки РПВ — так называемый план модифицированного обследования с применением гнездовой выборки (МОГВ) — был разработан в ответ на необходимость создания такой стратегии выборки, которая не предусматривала бы работ по составлению списков, но при этом базировалась бы на вероятностной выборке. Различные варианты плана МОГВ, а также другие планы были реализованы по всему миру в рамках проводимых под эгидой ЮНИСЕФ обследований по многим показателям с применением гнездовой выборки (ОМПГВ) для контроля выполнения определенных целей и задач Всемирной встречи на высшем уровне в интересах детей, касающихся положения детей и женщин (Международный фонд детского образования Организации Объединенных Наций, 2000 год).

147. План МОГВ является минималистской стратегией выборки. Он использует простую двух этапную выборку, предусматривающую тщательную стратификацию, а также быстрый обход района и его разделение на сегменты. Работы по составлению списка не проводятся. План выборки с помощью МОГВ имеет следующие основные характеристики:

- Формирование выборки территориальных единиц первого этапа, таких как деревни или городские микрорайоны, с применением выборки *с вероятностью, пропорциональной размеру*, или равновероятностной выборки в зависимости от уровня вариантности *ПЕВ* в отношении их показателей размера. Могут использоваться и старые показатели размера даже в том случае, если инструментарий выборки был сформирован уже несколько лет назад, но при этом такой инструментарий должен полностью охватывать обследуемую группу населения, как на национальном, так и на местном уровне.
- Посещение каждой территориальной единицы выборки с целью быстрого опроса населения, а также разбивка этой территории на сегменты с помощью существующих карт и картосхем, при этом число сегментов определяется заранее и равняется показателю размера согласно переписи, деленному на желаемый (ожидаемый) размер кластера. Создаваемые сегменты должны быть примерно равны по численности проживающего в них населения.
- Равновероятностный отбор одного территориального сегмента из каждой включенной в выборку ПЕВ.
- Проведение опросов всех домохозяйств в каждом отобранном сегменте.

148. Использование сегментации без составления списков является основным преимуществом плана согласно показателю размера. Он отличается от плана стандартных сегментов Обследования в области народонаселения и здравоохранения, который требует составления списка каждого сегмента. Работы по сегментации также частично вносят соответствующую поправку в случае использования инструментария, который может быть устаревшим. Хотя этот план имеет преимущество получения несмещенного результата, у него есть недостаток в плане более слабого контроля над конечным размером выборки, поскольку любой отобранный сегмент из-за роста населения может иметь гораздо более крупный размер, чем указано в инструментарии.

149. При этом, однако, составление карт требуется для плана согласно показателю размера точно так же, как и для плана стандартных сегментов Обследования в области народонаселения и здравоохранения со всеми присущими этой операции недостатками, которые отмечены в методе ОНЗ. Кроме того, сложности могут возникнуть с созданием небольших сегментов, совпадающих по своему размеру с размером кластера, с четко обозначенными границами в случаях, когда таких естественных границ для малых территорий просто не существует. Есть и последний недостаток: в той степени, в которой тот или иной сегмент, где проводится опрос, является компактным кластером, т. е. все домохозяйства в нем территориально прилегают друг к другу, это дает более сильный эффект схемы в силу сравнительно высокой внутриклассовой корреляции, чем эффект, вызываемый некомпактными кластерами плана стандартных сегментов.

3.9. Специальные темы: двухэтапные виды выборки и формирование выборки для определения трендов

150. Данный раздел охватывает две специальные темы, связанные с составлением плана выборки для обследований домашних хозяйств: *a*) двухэтапная выборка, в рамках которой первый этап используется для краткого опроса в целях выявления среди членов домохозяйств лиц, относящихся к обследуемой группе населения, а второй этап предусматривает формирование выборки из тех лиц, которые отвечают определенным критериям; и *b*) методология формирования выборки, согласно которой обследование проводится повторно для оценки изменений или трендов.

3.9.1. Двухэтапная выборка

151. Для обследований домашних хозяйств необходим особый вид плана выборки в том случае, когда не имеется достаточной информации для эффективного формирования выборки интересующей обследуемой совокупности. Такая необходимость, как правило, возникает, когда обследуемая совокупность представляет собой подгруппу населения — зачастую достаточно редкую, — члены которой присутствуют лишь в небольшой доле домохозяйств. Примерами могут служить лица определенной этнической группы, сироты или лица с доходом выше или ниже оговоренного уровня. Часто имеется возможность использовать тщательную стратификацию для выявления, например, территориальных единиц, где сконцентрированы интересующая этническая группа или лица с высоким доходом; однако, когда такие группы рассредоточены в относительно случайном порядке среди всего населения или когда какая-либо целевая группа — например, сироты — является редко встречающейся, тогда стратификация становится недостаточно эффективной стратегией, и должны применяться другие методы формирования выборки.

152. Одним из часто используемых методов является двухэтапная выборка, известная также как послестратификационная выборка или двойная выборка. Она осуществляется в четыре шага:

a) формирование «крупной» выборки домохозяйств;

b) проведение краткого проверочного опроса с целью выявления домохозяйств, члены которых относятся к обследуемой группе населения;

c) последующая стратификация крупной выборки на две категории по результатам проверочного опроса;

d) формирование подвыборки домохозяйств по каждой из двух страт для проведения второго, более подробного опроса целевой группы населения.

153. Целью такого двухступенчатого подхода является экономия затрат за счет проведения краткого проверочного опроса на этапе формирования первоначальной крупной выборки. За ним позднее следует более подробный опрос только в соответствующих критериям домашних хозяйствах. По этой причине в качестве первоначальной выборки зачастую выступает выборка, сформированная для иных целей, а проверочный опрос делается в качестве «довеска» к опросу в рамках основного обследования. Таким образом, эта процедура позволяет выделить основные ресурсы на проведение второго этапа выборки, и провести опрос на проверочном этапе в рамках весьма скромного бюджета.

Пример

Предположим, что планируется обследование 800 детей-сирот, проживающих в домохозяйствах родителей, здравствующих на момент опроса, или других родственников (в отличие от сирот, проживающих в специализированных учреждениях). Предположим далее, что согласно имеющимся оценкам для выявления 800 сирот необходимо включить в выборку 16 000 домохозяйств, т. е. 1 сирота на каждые 20 домохозяйств. Поскольку в связи со значительными расходами считается нецелесообразным формирование и опрос выборки из 16 000 домохозяйств ради проведения всего лишь 800 подробных опросов, принято решение использовать обследование общего назначения в области здравоохранения, которое также запланировано. Для обследования по вопросам здравоохранения сформирована выборка в 20 000 домохозяйств. Руководители обоих обследований договорились о том, что в вопросник обследования в области здравоохранения будет добавлен один дополнительный вопрос: «проживают ли в данном домохозяйстве лица в возрасте 17 лет или младше, у

которых отец, мать или оба родителя умерли?». Такой дополнительный вопрос, как ожидается, позволит выявить домохозяйства, в которых проживают примерно 1000 сирот. Руководитель обследования по проблемам сирот затем сформирует подвыборку из 80 процентов таких домохозяйств для проведения подробного опроса.

154. Приведенный выше пример также показывает, *в каких случаях* двухступенчатая выборка является подходящей стратегией. Следует отметить, что обследуемый размер выборки в данном примере составляет лишь 800 сирот, но при этом размер выборки с точки зрения числа домохозяйств, необходимых для выявления нужного числа сирот, составляет 16 000. Следовательно, при расчете последнего числа (см. формулу 3.7) технический специалист по формированию выборки и руководитель обследования скорее всего придут к выводу о том, что двухэтапная выборка представляется наиболее практическим и эффективным с точки зрения затрат планом выборки.

155. Последующая стратификация выборки первого этапа важна по двум причинам. Проверочный вопрос или вопросы практически всегда должны быть краткими, поскольку они прилагаются к другому обследованию, который и без того вероятнее всего предусматривает продолжительный опрос. Руководитель основного обследования вряд ли согласится на развернутый набор проверочных вопросов. Следовательно, существует вероятность того, что в ряде домохозяйств из упомянутого выше примера, в которых были выявлены сироты, не будет детей-сирот — и наоборот. Такие ошибки неправильной классификации требуют формирования двух страт: одной — для домохозяйств с положительным результатом проверки, другой — с отрицательным результатом. Выборки для проведения полного опроса будут формироваться из каждой страты, исходя из возможности наличия неправильной классификации. Доля выборки для «положительной» страты должна быть очень высокой — до 100 процентов, в то время как для «отрицательной» страты подойдет гораздо более низкая доля выборки.

3.9.2. Выборка для оценки изменений или трендов

156. Во многих странах обследования домашних хозяйств планируются с двойной целью оценки *a*) *базовых* показателей (их *уровней*) при первоначальном проведении обследования и *b*) *изменений* этих показателей в ходе второго и последующих обследований. Когда обследование повторяется более одного раза, измеряются также тренды изменения показателей. Проведение повторных обследований оказывает различные виды воздействий на план выборки, которые не проявляются при проведении единовременного многопрофильного обследования. В частности, проблемами, вызывающими наибольшую озабоченность, являются: надежность оценки изменений и определение надлежащей комбинации в плане использования одних и тех же или разных домохозяйств от одного обследования к следующему. С последним аспектом связаны опасения по поводу субъективности ответов и нагрузки на респондентов, когда из раза в раз проводится опрос одних и тех же домохозяйств.

157. Рассмотрение проблемы надежности также требует математической иллюстрации. Мы начнем с рассмотрения вариантности оцениваемых изменений, $d = p_1 - p_2$, которые выражаются следующей формулой:

$$\sigma_d^2 = \sigma_{p1}^2 + \sigma_{p2}^2 - 2\sigma_{p1,p2} = \sigma_{p1}^2 + \sigma_{p2}^2 - 2\rho\sigma_{p1}\sigma_{p2} \ , \qquad (3.22)$$

где величина p — это оцениваемая доля; σ_d^2 — вариантность разности; σ_p^2 — вариантность величины p при первом или втором раунде обследовании, которые обозначены цифрами 1 или 2; $\sigma_{p1,p2}$ —

ковариантность между *p1* и *p2; и* ρ — корреляция между полученными величинами *p1* и *p2* в ходе двух раундов обследования.

В тех случаях, когда оценка изменений незначительна, что случается достаточно часто, мы имеем

$$\sigma_{p1}^2 \approx \sigma_{p2}^2 ,$$

тогда, $\sigma_d^2 = 2\sigma_p^2 - 2\rho\sigma_p^2$ (мы можем опустить нижние индексы 1 и 2). Следовательно,

$$\sigma_d^2 = 2\sigma_p^2 (1-\rho). \qquad (3.23)$$

158. Для оценки уравнения (3.22) следует отметить, что оценка σ_p^2 для обследования с помощью гнездовой выборки является оценкой простой случайной выборки, ПСВ, умноженной на величину эффекта схемы, *deff.* Корреляция ρ, которая имеет максимальную величину при использовании той же выборки домохозяйств, может достигать 0,8 или даже больше. В этом случае оценка s_d^2 величины σ_d^2 рассчитывается как:

$$s_d^2 = 2\left[(pq) f/n\right](0,2) \quad \text{или} \quad 0,4\,(pq) f/n\,. \qquad (3.24)$$

159. При использовании одних и тех же кластеров, но других домохозяйств величина ρ все еще положительная, но значительно меньшая — вероятно в интервале от 0,25 до 0,35. Тогда мы имеем следующий результат (для ρ на уровне 0,3):

$$s_d^2 = 2\left[(pq) f/n\right](0,7) \quad \text{или} \quad 1,4\,(pq) f/n\,. \qquad (3.25)$$

160. И, наконец, при полностью независимой выборке во втором раунде обследования при применении различных кластеров и различных домохозяйств ρ равняется нулю, и мы имеем:

$$s_d^2 = 2\left[(pq) f/n\right]. \qquad (3.26)$$

При использовании типичного значения *deff* на уровне 2,0 формула 3.19 дает следующий результат:

$$s_d^2 = 4\left[(pq)/n\right]. \qquad (3.27)$$

161. Для повторных обследований с использованием частичного наложения выборки, например 50 процентов одних и тех же кластеров/домохозяйств и 50 процентов — новых, величина ρ должна быть помножена на коэффициент *F*, равный доле совпадающей части выборки. В этом случае уравнение 3.16 приобретает следующий вид:

$$\sigma_d^2 = 2\sigma_p^2 (1-F\rho). \qquad (3.28)$$

162. На основе указанных выше результатов можно сделать интересные выводы. Во-первых, оценка вариантности для сравнительно небольшого предполагаемого изменения между двумя обследованиями с использованием одной той же выборки домохозяйств составляет всего около 40 процентов от уровня вариантности, как по первому, так и по второму раунду обследования. Использование тех же кластеров, но других домохозяйств дает оценку вариантности на 40 процентов *выше* первого указанного уровня. Независимые выборки дают оценку вариантности *в два раза выше* этого уровня.

163. Следовательно, вариант с использованием одних и тех же домохозяйств при повторных обследованиях обладает существенными преимуществами с точки зрения надежности данных. При несоблюдении этого условия весьма значительное улучшение надежности достигается все же в случае использования либо *a*) доли одних и тех же домохозяйств, либо *b*) тех же кластеров, но с другими домохозяйствами. Обе стратегии дают результаты с более низкой вариантностью по сравнению с наименее удачным вариантом использования полностью независимых выборок.

164. Что касается проблемы ошибки регистрации данных, то наблюдается тем больше случаев возникновения двух негативных эффектов, связанных с респондентами — неполучение ответов и условный ответ, — чем чаще повторно используется одна и та же выборка домохозяйств. Респонденты не только все более неохотно принимают участие в опросах, увеличивая таким образом долю неполучения ответов в более поздних раундах обследования, но они также находятся под влиянием эффекта обусловливания, из-за чего при повторных опросах может снизиться качество или точность их ответов.

165. С указанным эффектом обусловливания связан феномен, известный как погрешность, вызванная «временем пребывания в выборке», когда ожидаются различные по величине оценки обследования от респондентов, ответы которых охватывают один и тот же период времени, но с разной степенью участия в данном обследовании. Этот феномен широко изучался, и было доказано его присутствие в обследованиях на самые различные темы: трудовые ресурсы, расходы, доходы и криминальная виктимизация. В Соединенных Штатах Америки, например, когда при проведении обследования рабочей силы респонденты опрашивались восемь раз, оценка уровня безработицы по ответам респондентов, в первых раз попавших в выборку, неизменно была примерно на семь процентов выше, чем в среднем по ответам респондентов за период проведения всех восьми опросов. Эта закономерность сохранялась в Соединенных Штатах в течение ряда лет. Анализируя эту погрешность, эксперты предположили, в том числе, что:

- регистраторы при последующих опросах могут не предоставлять респондентам такие же стимулы, как при первом опросе;
- респонденты могут узнать, что некоторые ответы вызывают дополнительные вопросы, и поэтому избегают давать определенные ответы;
- первый опрос может охватывать события, происшедшие за пределами учетного периода, в то время как в последующих опросах то или иное событие «привязано» к определенному сроку;
- респонденты действительно могут изменить свое поведение по причине участия в обследовании;
- респонденты могут быть не столь внимательны в плане предоставления точных ответов в последующих опросах, когда им наскучит сам процесс участия в обследовании (Kasprzyk, 1989).

166. Следует отметить, что большинство из указанных выше причин применимо в основном к повторным опросам для одного и того же обследования; но, тем не менее, некоторые проявления такого же поведения респондентов могут иметь место и в случаях, когда одни и те же домохозяйства используются для различных обследований.

167. Из вышеизложенного можно сделать вывод, что существуют конкурирующие друг с другом эффекты при использовании:

a) одной и той же выборки домохозяйств в каждом раунде обследования;

b) замены домохозяйств для части выборки;

c) новой выборки домохозяйств при каждом проведении обследования.

168. Если двигаться от *a* до *c*, то ошибка выборки по оценкам изменений увеличивается, в то время как ошибка регистрации данных имеет тенденцию к снижению. Ошибка выборки является наименьшей, когда в каждом случае используется одна и та же выборка домохозяйств в силу максимального коэффициента корреляции между измерениями. Противоположный эффект имеет место при использовании в каждом случае новой выборки домохозяйств.

169. Именно вариант *b* обычно рассматривается как наиболее компромиссный с точки зрения баланса ошибки выборки и погрешностей при регистрации данных. Если часть выборки сохраняется из года в год, ошибка выборки снижается по сравнению с вариантом *c*, а ошибка регистрации снижается по сравнению с вариантом *a*. Когда обследование проводится только в два раунда, наилучшим вариантом, по всей видимости, является вариант *a*. Воздействие респондентов, вероятно, не будет иметь слишком серьезных последствий для общего уровня ошибки обследования, когда выборка используется лишь дважды. Тем не менее наилучшим вариантом при повторном обследовании в три и более раундов будет вариант *b*. Удобной стратегией является замена в каждом раунде 50 процентов выборки по принципу ротации (примеры ротации выборки в эталонных выборках приведены в главе 4).

3.10. Если осуществление плана выборки нарушается

170. В настоящем разделе дается краткий обзор действий, которые необходимо предпринять в случае, если осуществление плана выборки сталкивается с какими-либо затруднениями, большинство из которых уже рассматривались или упоминались выше. При этом одним из важнейших принципов, который подчеркивается в этой и следующей главах, является то, что путем тщательного планирования во время составления плана выборки можно избежать многих препятствий в его осуществлении. Тем не менее, несмотря на самое тщательное планирование, могут возникнуть и непредвиденные проблемы.

3.10.1. Определение и охват обследуемого населения

171. Проблемы зачастую возникают по любой из множества причин, когда фактическая группа населения, охватываемая обследованием, не является его запланированной целевой совокупностью.

Пример

Рассмотрим обследование, в котором предполагается охватить типовую целевую совокупность в виде всего народонаселения страны. Фактически численность охватываемого насе-

ления (т. е. то, из которого формируется выборка) зачастую меньше, чем общая численность населения по любой из следующих причин:

- в выборку не входят лица, проживающие в специализированных учреждениях, таких как больницы, тюрьмы и военные казармы;
- лица, проживающие в определенных географических районах, могут целенаправленно исключаться из охвата. К таким районам могут относиться труднодоступные местности; территории, пострадавшие от природных бедствий; объявленные закрытыми из-за гражданских волнений или войны; поселки или лагеря, где живут беженцы и иностранные рабочие, и т. д.;
- лица, не имеющие постоянного места жительства, рассматриваются как "не входящие в сферу охвата" обследования. К ним могут относиться кочевое население, команды судов, временные работники и т. д.

172. Проблемы, касающиеся таких подгрупп населения, применительно к плану выборки связаны с тем, что они обычно перед проведением обследования не определяются как группы, подлежащие исключению из сферы охвата обследования. В связи с этим осуществление плана выборки нарушается, когда случайно выбирается, скажем, *a*) какой-либо кластер, который вместо «традиционного» жилого района оказывается трудовым лагерем, тюрьмой или общежитием, или *b*) какая-либо ПЕВ расположена в горной местности и считается недоступной. Часто принимаемым в таких ситуациях «решением» является замена на другую ПЕВ. Такое решение, однако, означает появление погрешности в процедуре выборки.

173. Приемлемое решение заключается в том, чтобы избежать такой проблемы на этапе составления плана выборки. Это достигается, во-первых, за счет тщательного определения обследуемой группы населения и определения не только входящих в нее подгрупп, но и тех, которые должны исключаться из охвата. Во-вторых, инструментарий выборки должен быть затем модифицирован с целью удаления из него любых территориальных единиц, не входящих в сферу охвата обследования. Это относится и к исключению любых специально созданных счетных участков — например, трудового лагеря рабочих. В-третьих, выборка должна формироваться из такого модифицированного инструментария. Более подробно инструментарий выборки рассматривается в главе 4.

174. Следует также иметь в виду, что предложенное выше решение позволяет более точно определить целевую совокупность. Важно также дать описание точной целевой совокупности в отчетах обследования для надлежащего информирования пользователей.

3.10.2. Размер выборки слишком велик для бюджета обследования

175. Другая проблема возникает, когда расчетный размер выборки оказывается более крупным, чем может обеспечить бюджет обследования. Когда это происходит, группа по проведению обследования должна либо изыскать дополнительное финансирование обследования, или модифицировать цели обследования в части подлежащих измерению показателей, уменьшая либо требования по точности данных, либо число областей обследования.

176. Одним из путей уменьшения точности (повышения ошибки выборки) в целях значительного снижения затрат является отбор меньшего числа *ПЕВ* при сохранении, однако, общего размера выборки. Например, вместо 600 ПЕВ, состоящих из 15 домохозяйств каждая (n = 9 000), план выборки может быть изменен в сторону отбора 400 ПЕВ из 22 или 23 домохозяйств каждая ($n \approx$

9 000). Что касается областей обследования, одно из решений может заключаться в том, чтобы ограничиться четырьмя основными регионами страны вместо, скажем, 10 провинций.

3.10.3. Размер кластера больше или меньше ожидаемого

177. Часто возникающая проблема состоит в том, что кластер выборки может значительно превышать свой показатель размера в результате, например, нового жилищного строительства, особенно в случае устаревшего инструментария выборки. Группа по проведению обследования может ожидать наличия 125 домохозяйств в заданном кластере, однако на этапе составления списка может обнаружить, что их количество составляет 400. Приемлемым решением в этом случае является разбивка кластера на территориальные подсегменты примерно равного размера с точки зрения их населения. Число сегментов должно равняться текущему количеству домохозяйств, деленному на первоначальный показатель размера, округленное до ближайшей целой величины. В нашем примере это будет значение 400/125 или 3,2, округленное до 3 сегментов. Эти сегменты должны создаваться путем нанесения их на карту и быстрого подсчета жилищных единиц (в отличие от домохозяйств). Затем для составления списка путем случайного выбора должен быть отобран один из сегментов.

178. Может также возникнуть и противоположная проблема. Кластер может оказаться значительно меньше ожидаемого в связи со сносом жилья, природным бедствием или по другим причинам. В этом случае часто возникает искушение заменить его на другой кластер, однако такой путь ведет к появлению погрешностей. Вместо этого меньший по размеру кластер должен быть использован таким, как он есть. Хотя это может привести к меньшему, чем целевой, конечному размеру выборки, возрастание ошибки выборки будет в конечном счете незначительным, если только не будет включено большое число таких кластеров. Использование меньшего по размеру кластера без изменения (или замены) позволит, тем не менее, получить несмещенную оценку, поскольку такой кластер является «репрезентативным» в плане текущих изменений его населения, происшедших с момента создания соответствующего инструментария.

3.10.4. Решение проблем, связанных со случаями неполучения ответов

179. Хотя эти случаи относятся скорее к проведению обследования, а не к осуществлению выборки, неполучение ответов представляет собой серьезную проблему, которая способна негативно повлиять на получение оценок обследования (проблема неполучения ответов подробно рассматривается в главах 6 и 8). Если допускается, что уровень неполучения ответов может превысить 10–15 процентов выборки, то возникающая в результате погрешность оценок может сделать их весьма сомнительными. И снова некоторые страны имеют тенденцию «решать» проблему неполучения ответов путем замены не дающих ответа домохозяйств. Такой метод сам по себе ведет к погрешностям, поскольку взятые на замену домохозяйства все же представляют собой только ответившие домохозяйства, но среди них нет неответивших. Известно, что характеристики последних двух групп по важным переменным показателям обследования должны отличаться друг от друга, особенно в части социально-экономического положения. Предпочтительным решением, которое, к сожалению, никогда не дает 100-процентного успеха, является получение ответов от первоначально не ответивших домохозяйств. Это должно достигаться с помощью планирования с самого начала возможности нескольких повторных посещений не ответивших домохозяйств с целью добиться их сотрудничества (в случае отказа) или застать их дома (при отсутствии по тем или иными причинам). Может понадобиться до пяти повторных посещений, однако минимальным должно быть три посещения.

3.11. Краткие рекомендации

180. В данном разделе дается краткое описание основных рекомендаций, которые необходимо вывести из этой главы. Хотя некоторые из рекомендаций применимы практически для любой ситуации (например, «использование вероятностной выборки»), есть другие рекомендации, для которых могут существовать исключения в зависимости от конкретных особенностей, ресурсов и требований той или иной страны. По этой причине руководящие указания, данные ниже в виде контрольного списка, представляются скорее в духе «эмпирических правил», а не как неизменные и непоколебимые рекомендации. Участники обследования должны стремиться:

- использовать методы вероятностной выборки на каждом этапе отбора;
- к максимальному упрощению, а не усложнению плана выборки;
- применять методы отбора, которые дают самовзвешенную или приблизительно самовзвешенную выборку в рамках областей обследования, а при отсутствии в плане таких областей — в рамках выборки в целом;
- использовать по возможности двухэтапный план выборки;
- рассчитывать размер выборки, используя формулу наподобие (3.5) и корректируя при необходимости величину постоянных параметров (таких, как прогнозируемая доля неполучения ответов и средний размер домохозяйства) с учетом специфики конкретной страны;
- использовать по умолчанию величину эффекта схемы на уровне 2,0 в формуле размера выборки, если для данной страны не имеется более точной информации;
- использовать в качестве основы размера выборки ту ключевую оценку, которая, как считается, будет охватывать наименьшую процентную долю совокупности среди всех ключевых оценок обследования;
- если позволяет бюджет, выбирать допустимый предел ошибки или уровень точности для ключевой оценки (см. выше) в размере 10 процентов от величины этой оценки, иными словами, относительная ошибка должна составлять 10 процентов *при 95-процентной доверительной вероятности*; или же добиваться уровня относительной ошибки в 12–15 процентов;
- выбирать первичные единицы выборки (ПЕВ) в качестве счетных участков переписи (СУ), если это удобно и возможно;
- использовать, где это возможно, неявную стратификацию совместно с систематической выборкой с вероятностью, пропорциональной размеру (ВПР), особенно в многоцелевых планах выборки;
- ограничивать до абсолютно необходимого минимума число областей проведения оценки (для доведения размера выборки до управляемого уровня);
- добиваться достаточно большого числа (нескольких сотен) кластеров (или ПЕВ при двухэтапной выборке): чем больше, тем лучше;
- использовать небольшие размеры кластеров (10–15 домохозяйств): чем меньше, тем лучше;

- использовать постоянный, а не переменный размер кластера, иными словами, фиксированное число домохозяйств вместо постоянной доли выборки;
- применительно к областям обследования добиваться минимального количества, равного 50 ПЕВ в каждой области;
- планировать минимум три, а предпочтительно — пять повторных посещений для решения проблемы неответивших домохозяйств;
- для редких подгрупп населения изучать возможность использования двухэтапной выборки путем включения "дополнительного" вопроса в уже запланированное крупное обследование в целях выявления обследуемых лиц и последующего подробного опроса в рамках подвыборки;
- применительно к обследованиям, предназначенным для измерения происшедших изменений, проводить опрос одних и тех же домохозяйств в обоих раундах только в том случае, если запланировано всего два опроса; при трех и более опросах использовать схему частичного совпадения выборки путем ротации новых домохозяйств в рамках выборки в каждом раунде.

Справочная литература и дополнительные источники информации

Bennett, S. (1993), The EPI cluster sampling method: a critical appraisal. Invited paper, International Statistical Institute Session, Florence, Italy.

Cochran, W. (1977), *Sampling Techniques,* 3rd ed., New York: Wiley.

Hansen, M., W. Hurwitz and W. Madow (1953). *Sample Survey Methods and Theory,* New York: Wiley.

Хуссманс Р., Мехран Ф., Верма В. *Обследования экономически активного населения: занятость, безработица и неполная занятость*. Методическое руководство Международного бюро труда. — М.: Финстатинформ, 1994. — В 2 т.

International Statistical Institute (1975). *Manual on Sample Design*. World Fertility Survey Basic Documentation. Voorburg, Netherlands.

Kalton, G. (1983), *Introduction to Survey Sampling.* Beverly Hills, California: Sage. Publications.

_____(1987), An assessment of the WHO Simplified Cluster Sampling Method for estimating immunization coverage. Report to UNICEF, New York.

_____(1993). *Sampling Rare and Elusive Populations. Department for Economic and Social Information and Policy Analysis,* Statistical Division and National Household Survey Capability Programme, INT-92-P80-16E. New York: United Nations.

Kasprzyk, D., and others, eds. (1989). *Panel Surveys.* New York: John Wiley & Sons, Chapter. 1.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Krewski, D., R. Platek, and J.N.K. Rao, eds. (1981). *Current Topics in Survey Sampling.* New York: Academic Press.

Le, T. and V. Verma (1997). *An Analysis of Sample Designs and Sampling Errors of the Demographic and Health Surveys*, DHS Analytical Reports, No. 3. Calverton, Maryland: Macro International Inc.

League of Arab States (1990), *Sampling Manual, Arab Maternal and Child Health Survey,* Basic Documentation 5. Cairo: Pan Arab Project for Child Development (PAPCHILD).

Macro International Inc. (1996). *Sampling Manual.* DHS-III Basic Documentation, No. 6. Calverton, Maryland: Macro International Inc.

Namboodiri, N., ed. (1978). *Survey Sampling and Measurement.* New York: Academic Press.

Raj, D. (1972). *Design of Sample Surveys,* New York: McGraw-Hill.

Scott, C. (1993), Discussant comments for session on "Inexpensive Survey Methods for Developing Countries". Invited paper, International Statistical Institute Session, Florence, Italy.

Som, R. (1966), *Practical Sampling Techniques,* 2nd ed., New York: Marcel Dekker, Inc.

Turner, A., R. Magnani, and M. Shuaib, (1996), A not quite as quick but much cleaner alternative to the Expanded Programme on Immunization (EPI) Cluster Survey design. *International Journal of Epidemiology,* (Liverpool, United Kingdom) Vol. 25, No. 1.

Организация Объединенных Наций (1984). *Руководство по обследованию домашних хозяйств* (Переработанное издание). Методологические исследования, № 31; в продаже под № R.83. XVII.13.

_____(1986). Sampling frames and sample designs for integrated household survey programmes, *National Household Survey Capability Programm:,* New York: United Nations. Department of Technical Cooperation for Development and Statistical Office.

_____ (2005). *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96. Организация Объединенных Наций, в продаже под № R.05.XVII.6.

United Nations Children's Fund (2000), *End-Decade Multiple Indicator Survey Manual.* Chapter. 4; entitled "Designing and selecting the sample", and appendix 7; entitled "*Sampling details*." New York: UNICEF.

United States Bureau of the Census (1978). *Current Population Survey Design and Methodology,* Technical Paper 40, Washington, D.C.: Bureau of the Census.

Verma, V. (1991), *Sampling Methods*. Training Handbook, Tokyo: Statistical Institute for Asia and the Pacific.

Waksberg, J. (1978), Sampling methods for random digit dialing, *Journal of the American Statistical Association,* vol. 73, pp. 40–46.

World Bank (1999). *Core Welfare Indicators Questionnaire (CWIQ) Handbook*. Washington D.C.: World Bank, chapter.4.

Health Organization (1991). *Expanded Programme on Immunization, Training for Mid-level Managers: Coverage Survey.* WHO/EPI/MLM91.10. Geneva.

Глава 4

Инструментарии выборки и эталонные выборки

4.1. Инструментарии выборки в обследованиях домашних хозяйств

1. В предыдущей главе были рассмотрены все многогранные аспекты плана выборки, за исключением инструментариев выборки, а также ряд вариантов составления плана выборки для обследований домашних хозяйств. Тем не менее одним из важнейших аспектов плана выборки обследований домашних хозяйств является ее инструментарий, и именно поэтому данному вопросу посвящается отдельная глава.

2. Инструментарий выборки оказывает значительное влияние на уровень затрат и качество любых обследований, касающихся как домашних хозяйств, так и других проблем. В обследованиях домашних хозяйств содержащие погрешности инструментарии выборки являются распространенным источником *ошибки регистрации данных*, в особенности неполного охвата важных подгрупп населения. В данной главе предпринята попытка подробно рассмотреть передовую практику составления и использования инструментария применительно к различным этапам выборки. Глава разделена на два раздела: первый из них охватывает общие вопросы инструментариев и их составления с упором на многоэтапный план выборки в обследованиях домашних хозяйств; второй — рассматривает вопросы специального характера, возникающие при использовании *эталонного* инструментария выборки.

4.1.1. Определение[1] инструментария выборки

3. Согласно простому рабочему определению инструментарий выборки — это *набор первоисточников, из которых формируется выборка*. Данное определение также раскрывает цель инструментариев выборки, а именно — предоставить средство для отбора конкретных лиц из целевой группы населения, которых предполагается опросить в ходе обследования. Для этого может потребоваться несколько наборов первоисточников. Это, как правило, относится к обследованиям домашних хозяйств в силу их многоэтапного формата. На ранних этапах отбора в обследованиях домашних хозяйств выборки обычно формируются из *территориальных* инструментариев. На последнем этапе выборки могут формироваться либо из территориального, либо из *списочного* инструментария (территориальный и списочный инструментарий рассматриваются ниже).

4.1.1.1. Инструментарий выборки и целевая совокупность

4. При принятии решения о применении надлежащего(их) инструментария(ев) в обследованиях домашних хозяйств важным соображением является взаимосвязь между целевой совокупно-

1 Читатель может обратиться к таблице 3.1 главы 3 для доступа к глоссарию связанных с выборкой терминов, которые используются в главах 3 и 4.

стью обследования и единицей отбора. Единица отбора определяет инструментарий. Она также определяет вероятность отбора на последнем этапе.

Пример

Возьмем в качестве иллюстрации обследование, целевой совокупностью которого являются грудные младенцы, причем группа по проведению обследования должна рассмотреть два возможных инструментария: в качестве одного могут выступать медицинские учреждения, которые вели регистрацию рождений в течение предыдущих 12 месяцев; в качестве другого — домохозяйства с младенцами в возрасте до 12 месяцев. В первом случае инструментарий будет состоять из двух частей (по одной на каждый этап отбора): во-первых, перечень больниц и клиник, в которых принимаются роды; во-вторых, список всех младенцев, родившихся в этих учреждениях за предыдущие 12 месяцев. Единицами отбора на первом этапе являются медицинские учреждения, на втором — младенцы. Следовательно, единица отбора и целевая совокупность совпадают в *конечном* этапе отбора. Во втором случае, однако, инструментарий по всей вероятности будет определяться (на последнем этапе отбора) как список домохозяйств по малым территориальным единицам, таким как деревни или городские микрорайоны. При использовании плана выборки будут проводиться отбор и проверка домохозяйств для подтверждения наличия в них детей в возрасте 0–12 месяцев. В этом случае домохозяйство будет выступать в качестве единицы отбора, на которой основывается вероятность отбора. Следует отметить, однако, что входящие в целевую совокупность лица фактически не выявляются и не обследуются вплоть за завершения проверки домохозяйств на предмет присутствия таких лиц. Следовательно, в случае инструментария, состоящего из списка домохозяйств, единица отбора и целевая совокупность отличаются друг от друга.

5. В обследованиях домашних хозяйств, т. е. в предмете данного руководства, единицей отбора, а также единицей, на которой базируется план выборки, является домашнее хозяйство. При этом целевая совокупность, даже в обследовании общего характера, будет иной в зависимости от подлежащих измерению целевых показателей. За исключением обследований доходов и расходов, целевая совокупность будет, как правило, представлять население, отличное от жителей самого домохозяйства. В качестве примеров можно привести: обследования занятости, в которых целевой совокупностью являются лица в возрасте 10 (или 14) лет и старше, что, таким образом, исключает всех детей младшего возраста; обследования репродуктивного здоровья женщин, целевой совокупностью которых являются женщины в возрасте 14–49 лет (а зачастую лишь женщины этой возрастной группы, когда-либо состоявшие в браке) и т. д.

4.1.2. Характеристики инструментариев выборки

6. Как указывалось выше, инструментарий выборки должен, безусловно, охватывать всю целевую совокупность со статистической точки зрения. Кроме того, идеальный инструментарий выборки должен быть *полным*, *точным* и *обновленным*. Это — идеальные характеристики, которые недостижимы в обследованиях домашних хозяйств. Тем не менее к ним необходимо стремиться, как при создании инструментария «с нуля», так и при использовании уже существующего инструментария. Качество инструментария может оцениваться с точки зрения того, насколько его идеализированные характеристики связаны с целевой совокупностью. Из главы 3 следует напомнить, что наше определение вероятностной выборки (т. е. той, в которой каждая единица целевой совокупности имеет известную, не равную нулю вероятность включения в выборку) выступает полезным «барометром» для оценки качества инструментария.

7. В зависимости от той степени, в которой не удалось достичь каждой из этих идеальных характеристик, результаты обследования будут иметь различные погрешности, но, чаще всего, в сторону *недооценки* целевой совокупности.

4.1.2.1. Полнота

8. Идеальный инструментарий считается полным в отношении целевой совокупности, если таким инструментарием охвачены все ее единицы (*генеральная совокупность*). Поэтому охват целевой(ых) совокупности(ей) является важнейшей характеристикой при оценке пригодности инструментария для того или иного обследования. Если он непригоден, то тогда группа по проведению обследования должна оценить, можно ли его исправить или доработать для обеспечения достаточной полноты. В предыдущем примере младенцы, родившиеся дома или в других местах вне медицинских учреждений, не попали бы в сферу охвата обследования при использовании медицинских учреждений в качестве единственного инструментария для формирования выборки. Следовательно, в этом примере значительная часть целевой совокупности будет иметь нулевую вероятность включения в выборку, т. е. нарушается условие формирования вероятностной выборки. В результате этого в инструментарии учреждений оценка числа младенцев будет заниженной. Более того, *характеристики* младенцев, вероятнее всего, будут существенно отличаться от характеристик младенцев, родившихся дома. Таким образом, инструментарий учреждений даст искаженное распределение важных показателей по младенцам и обеспечению ухода за ними.

9. Еще одной потенциальной проблемой при обследованиях домашних хозяйств является недостаточный охват. Например, план национального обследования может предназначаться для охвата всего населения с помощью обследования домашних хозяйств. Существуют, однако, различные подгруппы, такие как лица, находящиеся в специальных учреждениях, кочевые племена и команды судов, которые не проживают в домохозяйствах. В этом случае очевидно, что охват всего населения с помощью обследования домохозяйств невозможен. Необходимо создать дополнительные инструментарии для охвата не проживающих в домохозяйствах групп населения, для того чтобы такие лица имели отличную от нуля вероятность включения в выборку. При отсутствии таких мер возникает необходимость в модификации существующей целевой совокупности, с тем чтобы входящие в нее единицы поддавались более тщательному определению. В этом случае пользователям должна быть предоставлена четкая информация об исключенных из охвата обследованием групп населения.

4.1.2.2. Точность

10. Точность — это еще одна важная характеристика инструментария выборки, хотя неточности могут с большей вероятностью обнаруживаться в инструментариях, не применяемых в обследованиях домашних хозяйств. Можно утверждать, что инструментарий является точным, если в него один и только один раз включена каждая единица целевой совокупности. Рассмотрим список коммерческих предприятий с числом работников более 50 человек. Ошибки могут иметь место, если *a*) в список включено какое-либо предприятие с числом работников в 49 и менее человек, *b*) в списке отсутствует какое-либо предприятие с числом работников в 50 и более человек или *c*) какое-либо предприятие включено в список более одного раза (например, под разными названиями).

11. В обследованиях домашних хозяйств существует более низкая вероятность применения инструментариев, содержащих такие неточности. Однако они могут возникнуть в случаях, когда, например: *a*) в каком-либо инструментарии, состоящем из компьютерного файла счетных участков (СУ), отсутствуют некоторые его элементы, *b*) в списочном инструментарии домохозяйств

в какой-либо деревне отсутствуют некоторые домохозяйства, расположенные на окраинах этой деревни, *c*) в списочном инструментарии домохозяйств в ряде территориальных единиц некоторые домохозяйства включены в список более чем в одной такой единице или *d*) старый списочный инструментарий домохозяйств не включает недавно построенные жилые здания. Последний случай, а именно устаревший инструментарий, подробнее рассматривается ниже.

12. Факт отсутствия СУ или включенных в список домохозяйств в пределах территориальной единицы, безусловно, означает, что такие домохозяйства не имеют шанса быть включенными в выборку. Это опять же нарушает одно из условий формирования подлинно вероятностной выборки. Двойное включение в список также нарушает критерий вероятности, если только этот фактор не учитывается при расчете истинных значений вероятности отбора. К сожалению, указанные выше пропуски и ошибочное дублирование зачастую не выявляются. В связи с этим технический специалист по составлению выборки может и не знать о необходимости корректировки инструментария до формирования выборки на его основе. С другой стороны, небольшая доля случаев, когда имеют место выпавшие из инструментария или продублированные в нем единицы, обычно не вызывает какой-либо ощутимой — или даже заметной — погрешности в оценках из-за ошибки регистрации при обследовании.

4.1.2.3. Степень обновленности инструментария

13. Конечно, в идеальном случае инструментарий должен быть обновленным для соответствия двум другим критериям, а именно — полноте и точности. Устаревший инструментарий будет обязательно содержать неточности и с большой вероятностью будет неполным, особенно применительно к обследованиям домашних хозяйств. Типичным примером устаревшего инструментария является проводившаяся несколько лет назад перепись населения. Данные прошедшей переписи не могут точно отразить новое строительство или снос жилых зданий, лиц, переселяющихся в жилые помещения или выезжающих из них, случаи рождения или смерти. Эти недостатки не соответствуют требованию вероятностной выборки о том, что каждая единица целевой совокупности должна иметь *известную* вероятность включения в выборку.

Пример

Предположим, что инструментарий состоит из счетных участков, определенных согласно последней переписи населения, проведенной четыре года назад, причем обновление этого инструментария не проводилось. Предположим также, что на окраинах столицы образовались многочисленные районы самовольного заселения в тех *СУ*, которые на момент переписи были либо полностью, либо практически незаселенными. План выборки не предоставит домохозяйствам, проживающим в ранее пустовавших *СУ*, никакой возможности включения в выборку, нарушая, таким образом, условия вероятностной выборки. В тех *СУ*, которые были практически незаселенными, возникает другая серьезная проблема, даже несмотря на то, что такие *СУ* технически не нарушают требований вероятностной выборки. Выборка будет, несомненно, формироваться с использованием *вероятности, пропорциональной размеру*, и при этом *показателем размера* будут считаться данные переписи по численности населения или по домохозяйствам. В связи с весьма небольшой численностью населения во время переписи любой активно растущий *СУ* будет иметь очень малый шанс включения в выборку при использовании принципа формирования выборки с *вероятностью, пропорциональной размеру*. В итоге выборка будет иметь неприемлемо высокий показатель вариантности.

4.1.3. Территориальные инструментарии

14. В данном и следующем разделах мы обсудим две категории инструментариев, используемых при формировании выборки как в обследованиях домашних хозяйств, так и в других сферах применения. Важно отметить, что при многоэтапном плане выборки инструментарий для каждого этапа должен рассматриваться как отдельный элемент. Для каждого этапа используется свой собственный конкретный инструментарий. План выборки для обследования домашних хозяйств с большой вероятностью будет использовать и территориальный инструментарий (рассматриваемый в данном разделе) для первичных этапов, и списочный инструментарий (рассматриваемый в следующем разделе) для последнего этапа.

15. В обследованиях домашних хозяйств территориальный инструментарий выборки состоит из расположенных в иерархическом порядке территориальных единиц страны. Эти единицы в административном плане обозначаются по-разному в разных странах, однако, в нисходящем порядке обычно включают такие названия, как провинция или графство; округ; участок; административный район и деревня (в сельской местности) или (городской) микрорайон. Для целей переписи административные подразделения далее классифицируются на такие единицы, как директивные участки и счетные участки (*СУ*). Часто *счетные участки* переписи являются самыми мелкими определенными и имеющими четкие границы территориальными единицами страны.

16. Для целей обследований существуют четыре отдельные характеристики территориальных единиц, играющие важную роль для плана выборки:

a) они обычно полностью покрывают всю сухопутную территорию страны;

b) их границы четко определены;

c) для них имеются данные по численности населения;

d) они нанесены на карты.

17. Как отмечено выше, полный охват территории страны играет важную роль, поскольку это является одним из критериев достижения подлинно вероятностной выборки. Четко обозначенные и нанесенные на карту границы неоценимы при осуществлении плана выборки, так как они позволяют точно установить районы планируемого проведения работ на местах. Четкая информация о границах также помогает регистраторам найти вошедшие в выборку домохозяйства, отобранные в конечном счете для проведения опроса. Данные о численности населения необходимы в ходе составления плана выборки для присвоения показателей размера и расчета значений вероятности выборки.

18. Обычной точкой отсчета при создании территориального инструментария для обследований домашних хозяйств является перепись населения страны, исходя из приведенных выше четырех факторов. Кроме того, *счетный участок* — это имеющая удобный размер территориальная единица для целей отбора на последних этапах формирования выборки (предпоследний этап двухэтапного плана). В большинстве стран *СУ* целенаправленно формируются таким образом, чтобы включать примерно равное число домохозяйств — зачастую около 100 — для обеспечения сопоставимой рабочей нагрузки для регистраторов переписи.

19. Как ни парадоксально, территориальный инструментарий также представляет собой *список*, поскольку всегда необходимо начинать со списка административно-территориальных единиц совокупности для осуществления ранних этапов отбора при формировании выборки обследования домашних хозяйств. Для этого требуется рассмотреть списочные инструментарии.

4.1.4. Списочные инструментарии

20. Списочный инструментарий выборки — это, проще говоря, инструментарий, состоящий из списка единиц целевой совокупности. Теоретически списочный инструментарий для обследований домашних хозяйств существует в каждой стране сразу после проведения переписи. В принципе, недавняя перепись представляет географически упорядоченный список, который включает каждое домохозяйство или жилищную единицу страны.

21. Список только что проведенной переписи идеально подходит в качестве инструментария выборки домашних хозяйств, поскольку он является таким же обновленным, полным и точным, каким, в принципе, может быть любой список домохозяйств. Вследствие географической систематизации списка переписи, достаточно простой задачей является его стратификация в целях надлежащего географического распределения выборки. Следовательно, когда возникает необходимость в проведении выборочного обследования по итогам переписи для получения дополнительной информации или более подробной информации, нежели та, которую может эффективно дать перепись, список недавней переписи идеально подходит для использования в качестве списочного инструментария. Следует, однако, признать, что список недавней переписи только в течение короткого времени пригоден в качестве *действующего* инструментария. Совершенно очевидно, что чем длительнее период времени между переписью и последующим обследованием, тем менее применимыми становятся списки переписи в качестве источника для инструментария выборки.

22. Существуют и другие списки, которые в зависимости от их качества могут считаться надлежащими инструментариями выборки для обследований домашних хозяйств, например записи актов гражданского состояния и реестры подключения к коммунальным сетям. Регистры актов гражданского состояния могут быть возможными кандидатами для использования в качестве инструментария в тех странах, где ведется подробная отчетность, касающаяся граждан страны и их адресов. В некоторых случаях такие документы могут быть более полезными, чем территориальный инструментарий выборки, поскольку регистры, вероятнее всего, постоянно обновляются. Информация о подключении к коммунальным сетям — обычно к электроснабжению — может пригодиться в качестве инструментария выборки, если данные переписи серьезно устарели, однако сведения по коммунальным услугам должны быть оценены с точки зрения возможных проблем и их последствий. Очевидной проблемой, которая способная привести к неполному охвату, могут стать домохозяйства, не имеющие доступа к электроснабжению; другой проблемой, требующей решения, может стать использование системы подключения к электросети, которая обслуживает сразу несколько домохозяйств.

23. Другим видом списочного инструментария, который широко используется в развивающихся странах, является картотека телефонных абонентов. Формирование выборки осуществляется с помощью метода *случайного набора номера (СНН)* для обеспечения того, чтобы надлежащую возможность быть отобранными имели и абоненты, номера которых не включены в телефонный справочник. Однако формирование выборки методом случайного набора номера не рекомендуется для стран с низким уровнем охвата телефонных сетей.

24. В обычном обследовании домашних хозяйств последний этап отбора неизменно базируется на концепции списочного инструментария. Ранее мы рассмотрели, каким образом предпоследний этап плана выборки может дать гнездовую выборку, в рамках которой составляется текущий список домохозяйств. Из этого списка отбираются домохозяйства для включения в выборку. Таким образом, у нас имеются территориальный инструментарий, определяющий кластеры выборки, и списочный инструментарий, определяющий выборку домохозяйств в рамках этих кластеров.

4.1.5. Множественные инструментарии

25. В главе 3 был рассмотрен двухступенчатый план выборки для обследований домашних хозяйств. Он предусматривает применение методов проверки для выявления на первом этапе конкретной целевой группы населения, после чего на втором этапе проводится опрос подвыборки выявленных домохозяйств. Еще одним методом формирования выборки, дающим практически такой же конечный результат, является использование нескольких инструментариев выборки. Обычно это предполагает наличие только двух инструментариев, на основе которых составляется план по *двойному инструментарию*; однако иногда могут использоваться три и более инструментария (план по *множественному инструментарию*). Например, инструментарий совокупности, который определяется как сумма нескольких списков, из которых формируется независимая выборка, и в таком случае каждая часть множественного инструментария становится стратой (вопрос о стратификации рассматривается в разделе 3.4.2 главы 3 и в приложении I). Тем не менее общей проблемой для таких инструментариев остается дублирование данных.

4.1.5.1. Типовой двойной инструментарий в обследованиях домашних хозяйств

26. Для простоты изложения мы обсудим планы по двойному инструментарию, хотя общие принципы аналогичны в отношении планов по множественному инструментарию. Как правило, данная методология предполагает комбинирование территориального инструментария генеральной совокупности со списочным инструментарием тех лиц, о которых известно, что они входят в определенную обследуемую группу населения. Рассмотрим, например, обследование, целью которого является изучение данных о безработных лицах. Такое обследование может базироваться на территориальном инструментарии домохозяйств, но может быть дополнено выборкой на основе списочного инструментария безработных на текущий момент лиц, зарегистрированных в Министерстве социального обеспечения. Целью выборки такого типа, основанной на двойном инструментарии, является увеличение размера выборки за счет тех лиц, которые с большой вероятностью входят в обследуемую совокупность. Такой подход может стать более дешевой и более эффективной альтернативой двухступенчатой выборки. Инструментарий выборки общего назначения необходимо использовать для выявления не вошедших в список лиц из обследуемой группы населения. В данном примере к ним относятся безработные лица, не зарегистрированные в социальных службах.

27. Планы по двойному инструментарию, тем не менее, имеют ряд ограничений. Одним из них является то, что списочный инструментарий должен быть максимально обновленным. Если изменилось положение значительной части отобранных из списка лиц, что, по сути, вывело их из обследуемой группы населения, тогда использование списочного инструментария становится малоэффективным. В нашем примере тот факт, что трудоустройство безработного лица сделает его неподходящим по критериям выборки на момент проведения обследования, служит иллюстрацией того, почему списочный инструментарий должен быть максимально обновленным.

28. Другое ограничение обусловлено тем фактом, что места постоянного проживания вошедших в списочный инструментарий лиц могут быть рассредоточены по территории общины, что удорожит их опрос в связи с транспортными расходами. Это, конечно, резко контрастирует с территориальным инструментарием домохозяйств, из которого выборка может формироваться по кластерам для снижения расходов на проведение опроса.

29. Серьезной проблемой, связанной с планами по двойному инструментарию, является проблема дублирования данных. Как правило, лица, входящие в списочный инструментарий, будут также включены и в территориальный инструментарий. Возвращаясь к нашему примеру, без-

работные лица, отобранные из регистра, являются членами домохозяйств. Следовательно, при использовании обоих инструментариев у них будет удвоенный шанс включения в выборку. Проблему дублирования можно решить путем внесения надлежащих корректировок. Это, однако, повлияет на содержание вопросника обследования. В нашем примере каждому безработному, опрашиваемому в рамках выборки домохозяйств, следует задавать вопрос о том, зарегистрирован ли он/она в соответствующем министерстве в списке безработных. С ответившими положительно необходимо провести дальнейшую работу для сопоставления их имен со списочным инструментарием, причем этот процесс подвержен ошибкам и весьма сложен. При обнаружении совпадения должен быть изменен весовой коэффициент обследования в отношении данного лица на величину $(1/P_h + 1/P_l)$ для отражения того факта, что данное лицо имеет вероятность P_h отбора из инструментария домохозяйств и вероятность P_l отбора из списочного инструментария. Важно отметить, что такое сопоставление имен должно проводиться по всему списочному инструментарию, а не только в отношении вошедших в инструментарий лиц, которые также были включены в выборку. Это обусловлено тем, что вероятность (и вес) является функцией шанса отбора вне зависимости от того, произойдет ли фактически такой отбор.

4.1.5.2. Множественные инструментарии для различных видов жилых помещений

30. Еще один вид выборки по двойному инструментарию используется в случаях, когда обследуемое население проживает в различных видах жилых помещений, данные о которых не перекрываются. Например, обследование в отношение сирот, скорее всего, будет спланировано таким образом, чтобы включать сирот, проживающих в двух типах жилых помещений. Во-первых, учреждения типа сиротских приютов будут составлять один инструментарий. Во-вторых, придется включить в выборку и домохозяйства, чтобы охватить сирот, проживающих с одним из здравствующих на момент выборки родителей, другими родственниками или посторонними людьми. Следовательно, план по двойному инструментарию будет включать выборку из инструментария домохозяйств и выборку из инструментария учреждений, причем эти данные, естественно, не перекрываются.

31. Целью плана такого типа является адекватный охват обследуемого населения (по возможности, ближе к 100 процентам). Когда значительная часть населения проживает в одном из двух видов жилья, будут иметь место серьезные погрешности, если выборка будет ограничена только одним из инструментариев. Например, выборка сирот, сформированная только из тех, которые проживают в домохозяйствах, обусловит не только недооценку числа сирот, но и даст смещенную оценку с точки зрения касающихся их показателей. Аналогичные погрешности будут иметь место в случае, если обследование будет базироваться только на сиротах, проживающих в специальных учреждениях.

32. Рассмотренное выше ограничение в части дублирования данных не относится к планам выборки по двойным, не перекрывающимся друг с другом инструментариям. По этой причине их значительно легче осуществлять.

4.1.6. Типовые инструментарии для двухэтапных планов выборки

33. В главе 3 подчеркивалась практическая ценность двухэтапных планов выборки. В данном разделе обсуждается инструментарий, который обычно используется для двухэтапных планов.

34. В качестве территориальных единиц или кластеров, отбираемых на первом этапе отбора, зачастую используются деревни (или части деревень) или счетные участки переписи в сельских районах и микрорайоны в городах. В этом случае инструментарий состоит из всех территориальных единиц, составляющих обследуемую генеральную совокупность, каким бы образом она ни

была определена — страна в целом, провинция или ряд провинций, или столица. Формирование выборки осуществляется путем составления списка единиц, проверки его полноты, надлежащей стратификации списка (часто в территориальном отношении), а затем отбора (обычно с вероятностью, пропорциональной размеру) систематической выборки из данных единиц.

35. Если список кластеров генеральной совокупности слишком велик, могут вводиться промежуточные, фиктивные этапы отбора, которые были рассмотрены в предыдущей главе. В этом случае единицы инструментария определяются по-разному для каждого из фиктивных этапов. В приведенном ранее примере касательно Бангладеш единицами инструментария для двух фиктивных этапов выступали таны и союзы.

36. Единицами инструментария второго этапа в двухэтапном плане выборки выступают просто домохозяйства, находящиеся в кластерах первого этапа выборки. Когда они включаются в выборку из списка домохозяйств, то инструментарий по определению является списочным. Они могут также включаться в выборку в качестве компактных сегментов, созданных путем разбивки кластеров на самостоятельные и взаимоисключающие территориальные подразделения. В этом случае инструментарий второго этапа является территориальным.

4.1.7. Инструментарии эталонной выборки

37. В этом разделе мы дадим краткое описание понятия инструментария эталонной выборки, который более подробно рассматривается в разделе 4.2, ниже.

38. Инструментарий эталонной выборки — это инструментарий, используемый для формирования выборок либо для нескольких обследований, каждое из которых будет отличаться по содержанию, либо для использования в различных раундах постоянного или периодического обследования. За исключением периодического обновления по мере необходимости сам по себе инструментарий выборки не отличается ни для разных обследований, ни для отдельных раундов одного и того же обследования. Напротив — и это его отличительная характеристика — инструментарий эталонной выборки разрабатывается и составляется как стабильная, установившаяся система отбора подгрупп в выборке, которые необходимы для конкретных обследований или раундов одного и того же обследования в течение длительного периода времени.

4.1.8. Общие проблемы инструментариев и их предлагаемые решения

39. Проблемы, возникающие в обследованиях домашних хозяйств на основе дефектных инструментариев, включают как погрешности, не обусловленные выборкой, так и вариантность выборки. Как предполагалось выше, общие проблемы возникают тогда, когда инструментарий выборки является устаревшим, неточным или неполным. В подавляющем большинстве национальных обследований общего назначения в качестве базового инструментария выступает последняя перепись населения, и именно этот инструментарий рассматривается в данном разделе. Проблемы устаревания, неточности и неполноты в основанных на переписи инструментариях зачастую возникают одновременно. По мере увеличения интервала между переписью и обследованием масштаб этих проблем имеет тенденцию к увеличению.

40. Как уже упоминалось, инструментарий должен быть обновленным для того, что отражать текущие данные о состоянии населения. Инструментарий, базирующийся, скажем, на данных проводившейся пять лет назад переписи, не учитывает надлежащим образом рост и миграцию населения. Даже текущий, основанный на переписи инструментарий может быть неполным и вызвать проблемы при обследованиях домашних хозяйств, если он не охватывает лиц, прожи-

вающих в военных казармах, и команды судов, кочевые народы и другие значительные подгруппы населения, не проживающие в традиционных структурах домашних хозяйств. Неточности как текущих, так и старых инструментариев по данным переписи обусловливают самые разнообразные проблемы, включающие двойной учет домохозяйств в списках, домохозяйства, выпавшие из списков, или домохозяйства, зарегистрированные или закодированные по неправильным счетным участкам.

41. Надлежащие стратегии работы с устаревшими, неточными или неполными инструментариями, основанными на переписи, частично зависят от *a*) целей обследования и *b*) «возраста» самого инструментария. Что касается показателей измерения, если обследование целенаправленно планируется для охвата, например, только стационарных домохозяйств, то тогда подойдет и инструментарий на основе переписи, не учитывающий кочевые домохозяйства. С другой стороны, возникнет необходимость в разработке процедуры создания инструментария по кочевым домохозяйствам, если в задачи обследования входит охват таких домохозяйств (в странах, где они существуют). В этой связи вывод о том, является ли инструментарий полным или нет, зависит от определения обследуемого населения или обследуемых подгрупп населения.

42. Методы устранения проблем устаревания и неточностей инструментариев различаются в зависимости от срока, прошедшего с момента переписи. Хотя и неразумно было бы в связи с разнообразием условий той или иной страны предлагать какое-либо точное правило, эмпирическим правилом выбора надлежащей стратегии модернизации или обновления инструментария является определение того, превышает ли срок проведения последней переписи два года. Что касается неточностей того типа, которые упомянуты в пункте 40, то для их устранения применимы методы, которые приводятся в подразделах 4.1.2 и 4.1.8.

4.1.8.1. Инструментарии, основанные на переписи более чем двухлетней давности

43. Первая ситуация применима к странам с устаревшими данными переписи — двухлетней и более давности. Именно такие устаревшие инструментарии создают наиболее серьезные проблемы при составлении плана выборки обследований домашних хозяйств, особенно в быстро растущих городах. Идеальным решением является полное и проводимое в масштабах всей страны обновление инструментария, основанного на устаревших данных переписи, так как в случае успеха такого мероприятия, будет гарантия того, что получаемые в результате данные обследования будут максимально точными для целей надлежащего охвата и максимально надежными. К сожалению, такой путь является наиболее дорогостоящим, трудоемким и, следовательно, трудно реализуемым на практике. И, тем не менее, для стран с серьезно устаревшими данными переписи возможно не существует другой альтернативы.

44. Вместо полного обновления компромиссным решением будет обновление инструментария только по целевым районам, которые будут определяться экспертами той или иной страны, знакомыми с закономерностями роста и демографическими изменениями. Инструментарий, основанный на переписи, обновить достаточно просто: все, что нужно — это определить текущий показатель размера. Для целей обновления инструментария показатель размера должен устанавливаться как число жилых единиц в отличие от числа домохозяйств или лиц.

45. Важно осознать, что, для того чтобы методология выборки была действенной, нет необходимости в точном определении показателя размера. Например, если предполагается, что какой-то конкретный участок исходя из данных последней переписи включает 122 домохозяйства, то никаких поводов для беспокойства не возникнет, если его текущий размер составит 115 или 132 домохозяйства. По этой причине нецелесообразно пытаться обновлять инструментарий для старых,

сформировавшихся жилых районов, которые десятилетиями практически не меняются, несмотря на то что отдельные их жители могут приезжать или выезжать. Вместо этого, внимание должно привлекать кардинальное различие между текущей ситуацией и последней переписью — скажем, наличие 250 домохозяйств там, где ожидалось всего 100. Такие ситуации вероятны для жилых районов активного строительства или сноса жилья, таких как общины самовольного заселения на окраинах городов; районы многоэтажной жилой застройки или районы сноса жилых зданий. Именно такие районы должны стать целью обновления данных. Необходимо полагаться на представителей органов власти и экспертов той или иной страны для оказания помощи в выявлении таких целевых районов и, безусловно, проводить в работу только в тех районах, где изменения произошли уже после переписи.

46. Процесс обновления обычно включает несколько шагов, в том числе: *a*) выявление счетных участков, входящих в обследуемые районы, *b*) обход для быстрого подсчета выявленных счетных участков в целях получения текущего показателя размера и *c*) внесение изменений в файл переписи для отражения в нем обновленного показателя размера. Как уже упоминалось, достаточно иметь приблизительное значение показателя размера, и этим обусловлена необходимость проведения обхода для быстрого подсчета жилищных единиц, а не домохозяйств. При таком ускоренном обходе не обязательно в буквальном смысле стучать в двери для подсчета жилых помещений, за исключением, вероятно, многоквартирных зданий, где число жилищных единиц трудно установить, не входя в здание.

47. Обновление инструментариев на базе старой переписи описанным выше способом необходимо для стабилизации показателей вероятности включения в выборку единиц предпоследнего этапа, а следовательно, и надежности оценок обследования. В практическом плане обновление помогает держать под контролем не только общий размер выборки, но и рабочую нагрузку сотрудников на местах по составлению списков и проведению опросов. Более того, обновление снижает вероятность обнаружения на местах крупных кластеров, которые оказываются значительно больше ожидаемого размера, и необходимости их разбивки на подвыборки или принятия других соответствующих мер. С этим связан также тот факт, что любая подвыборка требует корректировки с помощью весовых коэффициентов, и это затрудняет процесс обработки данных. Потенциально можно снизить подобные возможности до такого уровня, когда после формирования выборки на предпоследнем этапе уже не будут обнаруживаться неожиданно крупные кластеры.

48. Стандартный план выборки для обследования домашних хозяйств будет, вероятнее всего, предусматривать составление текущего списка домохозяйств в кластерах выборки. В этом случае обновление должно осуществляться также и на предпоследнем этапе (см. ссылку в следующем разделе на применимую также и к данному случаю выборку с вероятностью, пропорциональной размеру). Таким образом, текущие списки кластеров выборки, которые не были обновлены, могут весьма близко напоминать списки переписи (хотя гарантий этого нет). Следует ожидать, однако, что кластеры выборки из обновленной части инструментария переписи дадут текущие списки, в значительной мере отличающиеся от списков переписи как по общему числу домохозяйств, так и по их конкретным адресам.

49. Необходимо подчеркнуть еще один важный момент касательно использования устаревших данных переписи, и он касается в большей мере достоверности выборки, а не ее вариантности. Как указывалось выше, кластеры обычно отбираются с вероятностью, пропорциональной размеру. Отсутствие обновления показателя размера для быстро растущих кластеров до начала формирования выборки приведет в результате к серьезной недопредставленности районов с небольшой численностью домохозяйств на момент переписи, но значительно выросших с того периода. Показатели обследования будут искаженными и, конечно, вводящими в заблуждение, поскольку

характеристики лиц, проживающих в таких быстро растущих районах, будут, вероятно, весьма сильно отличаться от характеристик лиц, проживающих в более стабильных жилых районах.

4.1.8.2. Инструментарии по данным переписи, проводившейся два и менее года назад

50. Данный раздел применяется в отношении тех стран, которые провели переписи сравнительно недавно — один или два года назад — и, следовательно, не нуждаются в общем обновлении инструментария. В таких случаях кластеры должны отбираться с применением первоначального действовавшего в период переписи *показателя размера*, поскольку он, как ожидается, будет весьма точным. Обновление, как таковое, должно проводиться только на предпоследнем этапе отбора после того, как сотрудники на местах составят текущие списки домохозяйств во включенных в выборку кластерах. Домохозяйства для включения в выборку должны отбираться из текущих списков, а веса выборки должны корректироваться, по мере необходимости, в соответствии с процедурами, рассмотренными в разделе 3.7.2 применительно к формированию выборки с вероятностью, пропорциональной оцениваемому размеру.

51. Хотя несколько кластеров в генеральной совокупности инструментария могли значительно вырасти с момента завершения переписи, предполагается, что число таких случаев будет не столь большим, чтобы существенно повлиять на проведение работы на местах или на точность оценок обследования. Любые такие кластеры, попавшие в выборку, могут при необходимости разбиваться на сегменты. Разбивка на сегменты или так называемая «фрагментация», как известно, представляет собой проводимую на местах операцию, направленную на снижение объема работ по составлению списков. Данная процедура включает *a*) разделение первоначального кластера на отдельные секции — обычно квадранты, *b*) случайный отбор одной секции для составления списка и *c*) отбор домохозяйств из данного сегмента для проведения опроса. Фрагментация не повышает уровень надежности выборки, так как каждому фрагменту выборки будет придан дополнительный весовой коэффициент обследования, равный числу фрагментов в кластере, — например коэффициент четыре, если кластер поделен на квадранты. Фрагментация, тем не менее, помогает сдерживать расходы по проведению работы на местах. Необходимость фрагментации может возникнуть даже при недавно состоявшейся переписи, и вновь это касается быстрорастущих счетных участков, которые претерпели радикальные изменения с момента переписи. Конечно, если перепись проводилась совсем недавно, таких районов будет очень мало.

52. Необходимо отметить, что упомянутые ранее виды погрешностей (учтенные дважды или отсутствующие в списке домохозяйства, неправильное распределение по счетным участкам) частично корректируются в ходе обновления, включающего составление новых списков домохозяйств на предпоследнем этапе. Это, естественно, может служить еще одной серьезной причиной составления обновленных списков домохозяйств в рамках обследований.

4.1.8.3. Случаи, когда инструментарий используется для других целей

53. Руководители обследований иногда ставят под вопрос возможность использования инструментария домохозяйств, составленного специально для одного вида обследования, в другом обследовании. Можно ли инструментарий выборки, предназначенный для обследования рабочей силы, использовать, например, в плане выборки для измерения состояния здоровья населения, показателей инвалидности, уровня бедности или фермерских хозяйств? Как правило, однако, проблемой является не сам инструментарий, а метод его стратификации. В большинстве случаев инструментарии могут использоваться в различных видах обследований, за исключением случаев, когда они являются неполными, неточными и устаревшими для того или иного планируемого

обследования. Ниже приводится пример инструментария, который представляется неадекватным для нескольких областей применения. Например, если обследование, основным вопросом которого является стоимость жизни, базируется лишь на городском населении (что часто происходит на практике), инструментарий выборки не будет содержать сельские районы. Ясно, что такой инструментарий не подойдет для оценки уровня бедности в странах, где это явление характерно в основном для сельских районов.

54. Большинство обследований домашних хозяйств, тем не менее, носят общий характер не только с точки зрения их содержания, но и с точки зрения их планов выборки. Обследование рабочей силы обычно включает, например, получение вспомогательной информации по демографическим характеристикам, уровню образования и по другим темам. В таких случаях соответствующий план выборки также носит общий характер, что предполагает использование традиционного инструментария выборки — а именно того, который охватывает все домохозяйства страны. Этот инструментарий может быть стратифицирован по той или иной переменной величине, специально предназначенной для измерений по тематике рабочей силы. Например, счетные участки могут быть классифицированы по такой переменной величине, как доля безработных согласно данным последней переписи. Затем в счетных участках могут быть созданы три страты — низкий, средний и высокий уровни безработицы. Как упоминалось выше, это решение касается только стратификации. Сам инструментарий остается неизменным. Решением проблемы может быть «дестратификация» инструментария, если его необходимо использовать для другого вида обследования, такого как обследование сферы здравоохранения.

55. Важнейшей задачей ответственного за выборку специалиста-статистика является оценка имеющегося инструментария выборки, когда этот инструментарий должен использоваться в другом виде обследования. Такая оценка предусматривает обеспечение того, чтобы инструментарий в его текущем виде соответствовал целям измерения показателей планируемого обследования, главным образом в части, касающейся полноты, точности и обновленности (что обсуждается в настоящей главе).

4.2. Инструментарий эталонной выборки

56. Эталонные выборки могут быть эффективными и экономичными с точки зрения затрат, чтобы выдержать их применение, когда та или иная страна проводит достаточно большое количество независимых друг от друга обследований или регулярных раундов одного и того же обследования. Само собой разумеется, что такие выборки должны быть надлежащим образом спланированы, но столь же важно поддерживать их в рабочем состоянии с течением времени. Организация Объединенных Наций (1986 год) предоставляет результаты гораздо более детальных исследований по инструментариям эталонной выборки и способам их использования.

4.2.1. Определение и использование эталонной выборки

57. Инструментарий (или инструментарии) выборки для первого этапа отбора в обследовании домашних хозяйств должен охватывать всю целевую совокупность. Когда инструментарий используется для многих обследований или нескольких раундов одного и того же обследования, он называется инструментарием эталонной выборки или просто эталонной выборкой.

58. Использование инструментария эталонной выборки является предпочтительной стратегией для любой страны, которая в период между переписями проводит крупномасштабную, постоянную программу обследований домашних хозяйств. И наоборот, при отсутствии постоянной

программы обследований использование эталонных выборок, как правило, не рекомендуется. При использовании одних и тех же единиц инструментария с течением времени возникает экономия масштаба, поскольку значительная часть расходов на формирование выборки приходится на операции по составлению эталонного инструментария выборки, а не на проведение работ на местах в рамках того или иного обследования. С другой стороны, те страны, которые проводят национальные обследования лишь эпизодически в период между переписями населения, не получат заметной выгоды от применения планов эталонной выборки.

59. К характеристикам эталонной выборки относятся число, размер и вид единиц на первом этапе отбора. Как правило, эталонная выборка состоит из начального отбора первичных единиц выборки (ПЕВ), которые остаются постоянными для каждой подвыборки. Следует отметить, что на более поздних этапах они обычно варьируются. Например, на последнем этапе отбора те конкретные домохозяйства, которые отбираются для опроса, обычно отличаются в независимых друг от друга обследованиях, но при этом они могут быть одними и теми же или частично перекрываться в повторных обследованиях.

4.2.2. Идеальные параметры первичных единиц выборки для инструментария эталонной выборки

60. Руководящие принципы составления инструментария эталонной выборки лишь незначительно отличаются от таких принципов в отношении инструментариев выборки в целом. Эталонный инструментарий должен быть настолько полным, точным и обновленным, насколько это практически осуществимо. Инструментарий эталонной выборки для обследований домашних хозяйств, так же как и обычный инструментарий выборки, как правило, составляется на основе данных последней переписи. Тем не менее, поскольку эталонный инструментарий может использоваться в течение всего периода между переписями, он обычно требует периодического и регулярного обновления, например каждые два-три года. Этот подход существенно отличается от обычного инструментария, который, вероятнее всего, будет обновляться внепланово и только тогда, когда планируется какое-либо определенное обследование.

61. Характеристики, способствующие составлению эталонного инструментария, вполне предсказуемо аналогичны характеристикам инструментариев выборки в целом. Определение единиц, которые должны использоваться в качестве первичных единиц выборки, ограничено, например, требованием о том, что они должны являться уже нанесенными на карту территориальными единицами. Однако это — не столь жесткое ограничение, поскольку в качестве единиц инструментария будут неминуемо выбраны административные единицы, созданные для целей переписи. Тем не менее важным требованием, которое может отличаться от требований к обычным инструментариям выборки, является то, что размер первичных единиц выборки должен быть достаточно большим для обслуживания многих обследований без необходимости регулярного опроса одних и тех же респондентов, однако в некоторых сферах применения это требование может быть ослаблено.

Пример

Определенный вид эталонного инструментария, который уже использовался в ряде случаев, базируется на двухэтапном плане выборки. Первый этап предусматривает крупную выборку *счетных участков* (или аналогичных небольших по размеру и нанесенных на карту территориальных единиц). Подвыборка из эталонной выборки *счетных участков* формируется для каждого независимого обследования, в котором используется данный инструментарий. По каждой подвыборке составляется список, или она иным образом фрагментируется для

использования в конкретном обследовании на момент активного планирования последнего. В качестве дополнительной иллюстрации эталонная выборка может состоять из 10 000 *счетных участков*, из которых 1 000 СУ входит в подвыборку для обследования занятости. В этой тысяче *счетных участков* составляются списки домохозяйств, из которых для обследования на втором этапе формируется выборка в 15 домохозяйств по каждому *счетному участку*. На следующий год из данной эталонной выборки формируется еще одна подвыборка из 800 *счетных участков* для использования в обследовании сферы здравоохранения и т. д. Таким образом, ни один из *счетных участков* не используется более одного раза, и, следовательно, размер *первичной единицы выборки* не имеет значения.

62. Однако размер *первичной единицы выборки* играет важную роль, когда все подвыборки, создаваемые на основе эталонного инструментария, должны формироваться из одного и того же набора *первичных единиц выборки*; в приведенном выше примере *первичными единицами выборки* являются *счетные участки*, при этом в каждой подвыборке используется различная группа *счетных участков*. Отбор *первичных единиц выборки* в эталонной выборке не является особой проблемой, поскольку как для эталонной, так и для любой другой выборки применяется один и тот же метод. Как правило, в качестве этого метода будет использоваться выборка с вероятностью, пропорциональной размеру (*впр*), кроме ряда редких случаев, когда *первичные единицы выборки* имеют более или менее равный размер, и в такой ситуации может использоваться равновероятностная выборка *первичных единиц выборки*.

4.2.3. Использование эталонной выборки для обеспечения проведения обследований

63. В разделе 3.3.7 обсуждался вопрос о том, что крупная выборка необходима при разработке эталонных выборок в целях предоставления достаточного числа домохозяйств для обеспечения проведения многочисленных обследований в течение нескольких лет без необходимости повторного опроса одних и тех же респондентов. Предполагаемые размеры выборки для всех планируемых и возможных обследований, в которых может применяться инструментарий эталонной выборки, являются ключевыми параметрами при создании его структуры. Например, если предполагается, что в различных обследованиях, использующих эталонную выборку, будет опрошено 50 000 домохозяйств, то группа составителей выборки будет располагать необходимой ей базовой информацией для принятия решения о числе и размере *первичных единиц выборки*. Более того, как показано в следующем примере, план проведения обследования можно разработать, исходя из использования эталонной выборки (для сравнения см. также пример в разделе 3.3.7).

Пример

Как и в примере из раздела 3.3.7, эталонная выборка для страны А состоит из 50 000 домохозяйств. Три запланированных обследования будут иметь следующие размеры выборки и кластеров: 16 000 домохозяйств, 6 домохозяйств на один кластер для обследования доходов и расходов; 12 000 домохозяйств, 12 домохозяйств на один кластер для обследования рабочей силы; 10 000 домохозяйств, 20 домохозяйств на один кластер для обследования по вопросам здравоохранения. Различные размеры кластеров выбраны для компенсации дифференциальных эффектов *deff* (в зависимости от вида обследования). Кроме того, 12 000 домохозяйств необходимо оставить в резерве для других обследований в случае необходимости. Три запланированных обследования потребуют в общей сложности 4 167 *первичных единиц выборки* (16 000/6 + 12 000/12 + 10 000/20). Поскольку неизвестно содержание обследований, в которых может использоваться резервная подвыборка, принято решение планировать размер кластера на уровне 12, что добавляет еще 1 000 *первичных единиц выборки* до итоговой

суммы в 5 167 единиц. Вследствие этого группа по составлению плана эталонной выборки решает сформировать эталонную выборку из 5 200 *первичных единиц выборки*. При определении *первичной единицы выборки* необходимо принимать во внимание число подлежащих опросу домохозяйств. В данном примере каждая *первичная единица выборки* должна быть достаточно большой для того, чтобы предоставить 50 домохозяйств для опроса. Имея эту информацию, группа по формированию выборки может затем определить, какая из территориальных единиц наилучшим образом соответствует определению *первичных единиц выборки*. Если в стране А имеются *счетные участки*, в которые в среднем входят 100 домохозяйств с небольшими отклонениями от этой средней величины, тогда имеет смысл использовать *счетные участки* в качестве *первичных единиц выборки*.

4.2.3.1. Преимущества многократного использования инструментария эталонной выборки

64. Существуют очевидные преимущества использования эталонной выборки. Первое и главное — план эталонной выборки служит в качестве инструмента координации для отраслевых министерств и других организаций, участвующих в любой национальной статистической программе. Это применимо к ряду аспектов проведения обследований, лежащих за пределами собственно формирования выборки, особенно в части контроля за расходами и разработки для различных секторов стандартных процедур в части определений статистических терминов, формулировок вопросов обследования и процедур кодирования данных.

65. Основным преимуществом программы эталонной выборки является использование одних и тех же *первичных единиц выборки*. Персонал на местах можно организовать и содержать в течение всего цикла использования эталонной выборки. Например, регистраторов можно нанять, подготовить и откомандировать к началу задействования программы эталонной выборки, когда будет известно, где расположены *первичные единицы выборки* для всех обследований, в которых предполагается использование этого инструментария в течение, скажем, 10 лет. В необходимом числе регистраторы могут наниматься на местах из числа лиц, проживающих в *первичных единицах эталонной выборки* или недалеко от них. Такие материалы обследования, как карты *первичных единиц выборки* и таблицы со списками домохозяйств, могут быть составлены в начале программы эталонной выборки, экономя, таким образом, время, а также списывая значительную долю организационных расходов для всех планируемых обследований. Кроме того, многократное использование такой выборки дает возможность более жесткого контроля ошибок регистрации данных и даже доли неполучения ответов. Это обусловлено тем, что повторные посещения одних и тех же респондентов позволят зафиксировать отношение таких респондентов к обследованиям и отметить проблемы в различных областях для принятия мер по их исправлению в последующих обследованиях. При этом необходимо вновь подчеркнуть, что такие преимущества появляются только в случае активного использования эталонной выборки.

66. Как правило, к прочим преимуществам эталонных выборок, как при использовании одних и тех же первичных единиц выборки в трехэтапном плане, так и разных первичных единиц выборки в двухэтапном плане, относится возможность *a*) в аналитических целях объединения данных из двух и более обследований различного содержания с применением эталонной выборки и *b*) быстрого реагирования на непредвиденные потребности в части сбора данных.

4.2.3.2. Недостатки многократного использования инструментария эталонной выборки

67. Методу эталонной выборки присущи определенные потенциальные недостатки, такие как полная выработка ресурса имеющихся первичных единиц выборки, иными словами, исчерпание

существующих домашних хозяйств при чрезмерном использовании эталонного инструментария. Это, однако, можно предвосхитить с помощью надлежащего предварительного планирования. Хотя и невозможно предвидеть все потенциальные случаи использования эталонной выборки на протяжении всего ее жизненного цикла, можно разработать резервные «подподвыборки» для их вероятного использования, если, конечно, сама эталонная выборка имеет достаточно большой размер.

68. Другим недостатком является постоянно возрастающий объем погрешностей, которые накапливаются при отсутствии надлежащего обновления по мере устаревания эталонной выборки. И наконец, эталонная выборки малопригодна для предоставления данных в обследованиях, проводимых согласно «особым требованиям», таким как обследования конкретных провинций или редких подгрупп населения, которые могут заинтересовать пользователей.

4.2.4. Распределение выборки между несколькими областями обследования (административные регионы и т. д.)

69. Национальные статистические службы испытывают на себе растущее давление в плане создания таблиц и анализа имеющихся у них данных обследований домашних хозяйств по важным административным единицам субнационального уровня, таким как крупные регионы, провинции и большие города. В некоторых странах, например во Вьетнаме, от статистических служб требуется регулярное предоставление данных даже на уровне районов. Такие требования и ожидания обусловлены вполне закономерными политическими потребностями, обычно в связи с тем, что социально-экономические программы предназначаются и разрабатываются для конкретных районов в отличие от страны в целом.

70. Выражаясь языком статистической выборки, эти данные называются оценками областей обследования, как уже упоминалось выше. Поскольку для получения надежных результатов необходимы огромные размеры выборки, это требует значительных средств, которые зачастую выходят за пределы бюджетных возможностей правительств на проведение обследований. Потребность в получении данных по областям обследования также влияет на составление эталонного инструментария выборки.

71. Число и вид необходимых для формирования выборки областей обследования рассматривались в разделе 3.3.4, посвященном размерам выборки, и этот вопрос здесь повторно затрагиваться не будет. После принятия соответствующих решений можно приступать к созданию инструментария эталонной выборки. Например, одна страна может принять решение о том, что обследования, проводимые на основе ее программы эталонной выборки, должны также предоставлять данные по двум изучаемым областям — городским и сельским районам. Другая страна, имеющая 12 провинций и желающая предоставить оценки по всем этим провинциям, может посчитать, что имеющиеся у нее ресурсы на проведение обследований способны обеспечить более крупные размеры выборки, которые необходимы в случае, если провинции будут рассматриваться в качестве областей обследования. Третья страна, в состав которой входят 50 провинций, может решить, что предоставление оценок по каждой из них требует слишком больших затрат. Вместо этого в такой стране может быть принято решение определять в качестве областей обследования 8 основных географических регионов, на которые разделяются 50 провинций. Четвертая страна может принять решение не вводить области, как таковые, а вместо этого составлять таблицы с помощью национальной выборки, пропорционально распределяемой по регионам, провинциям, городским и сельским районам и по отдельным крупным городам, предполагая при этом опубликовать данные по тем районам, по которым размер выборки будет оценен, как достаточно большой для получения обоснованно надежных результатов.

72. Применительно ко всем приведенным выше примерам распределение областей обследования неактуально, поскольку выборка распределяется пропорционально по представляющим интерес районам субнационального уровня. По первым трем примерам необходимо принять особые меры для надлежащего распределения первичных единиц эталонной выборки. Поскольку оценка результатов по областям обследования предполагает одинаковый уровень надежности для каждой подгруппы населения или территориальной единицы, определенной в качестве указанной области, для каждой области обследования должно отбираться одинаковое число первичных единиц выборки, причем это требование справедливо вне зависимости от того, базируется ли план выборки на эталонной выборке или нет.

4.2.5. Ведение и обновление эталонных выборок

73. С точки зрения влияния на охват населения надлежащее ведение и обновление инструментария эталонной выборки являются основным фактором при его составлении и планировании его использования. Эталонная выборка той или иной страны, как правило, используется в течение 10-летнего периода между переписями, на протяжении которого с большой вероятностью произойдут далекоидущие изменения в движении населения. Необходимо периодически обновлять инструментарий для отражения в нем демографических изменений, с тем чтобы такой инструментарий сохранял свою «репрезентативность».

74. Важно проводить два вида обновлений. Первый, более простой вид предусматривает составление новых списков домохозяйств в кластерах выборки, отобранных на предпоследнем этапе, причем данная процедура в общем рекомендуется для применения по всему тексту руководства как для эталонных выборок, так и для планов выборки одноразового использования. Таким образом, кластеры выборки автоматически обновляются для отражения в них миграции населения, рождений и случаев смерти. Этот вид обновления, при его ограничении только кластерами выборки, помогает свести к минимуму ошибку неполного охвата (ошибку регистрации данных), при этом, однако, если не обновлять весь инструментарий в целом, то со временем будет возрастать вариантность выборки.

75. Существует также необходимость в периодическом обновлении всего инструментария, с тем чтобы в нем надлежащим образом учитывался происходящий после переписи рост населения в широких масштабах. Как обсуждалось выше, такой рост имеет место в случае многоэтажного жилищного строительства и расширения районов самовольного заселения в городах. Те счетные участки, в которых наблюдается такой рост, на момент составления инструментария эталонной выборки были, безусловно, значительно меньше. В результате их параметры размера по мере роста населения становятся существенно заниженными. Следовательно, сводится к минимуму шанс их отбора при использовании выборки с вероятностью, пропорциональной размеру. Выбор таких счетных участков способен радикально повлиять на уровень вариантности выборки, поскольку их текущий показатель размера может быть на порядок больше первоначальной величины.

76. Существует возможность значительного смягчения проблем быстрорастущих районов и их вредного влияния на эталонную выборку путем регулярной проверки инструментария, скажем, каждые два-три года.

4.2.6. Ротация первичных единиц выборки в эталонных выборках

77. Читатель может обратиться к разделу 3.9.2 главы 3 под названием «Выборка для оценки изменений или трендов», где подробно рассматривается проблема частичного совпадения выборок в связи с периодическими или непрерывными обследованиями, предназначенными для измерения

изменений или трендов. Частичное совпадение выборок предполагает наличие схемы выборки, которая предусматривает замещение одних домохозяйств другими при повторном обследовании. Необходимо еще раз подчеркнуть, что использование частично совпадающих выборок является предпочтительным методом для оценки изменений, скажем, на годовом уровне. Одним из методов замещения является ротация выборки, которая обеспечивает частичное совпадение выборок между отдельными раундами обследования.

78. В предыдущем разделе отмечалось, что как надежность выборки (что желательно), так и ошибка регистрации данных (что нежелательно) имеет наибольшие значения в случае использования одних и тех же домохозяйств в каждом раунде обследования. Поэтому обычно находится компромиссное решение путем частичного совпадения выборок от одного раунда обследования к другому, особенно когда такое обследование повторяется три и более раз (обоснование применения частичного совпадения выборок дается в разделе 3.9.2).

79. Одним из методов введения частичного совпадения является замена или ротация первичных единиц выборки (в отличие от замены домохозяйств в рамках первичных единиц выборки). Одинаково важно внимательно изучить потребность в ротации ПЕВ, когда эталонная выборка первичных единиц выборки используется не только для отдельных раундов одного и того же обследования, но и для многих других обследований.

80. Для того чтобы план ротации был не только реализован на практике, но и дал значимые результаты, степень перекрытия между отдельными периодами времени должна быть одинаковой и постоянной во времени. Например, если доля совпадаемых единиц между первым и вторым годом составляет К процентов, тогда она должна быть К процентов и для периода между вторым и третьим годом, между третьим и четвертым годом и т. д. Соответственно, когда проводится полная ротация первичных единиц выборки, этот фактор должен быть учтен в плане ротации.

4.2.6.1. Примеры эталонных выборок для отдельных стран

81. В данном разделе даются описания эталонных выборок в четырех развивающихся странах: Камбодже, Объединенных Арабских Эмиратах, Вьетнаме и Мозамбике. Каждое из описаний иллюстрирует характеристики и принципы формирования эталонной выборки, такие как одно- или двухэтапная выборка эталонных единиц выборки и гибкое применение эталонной выборки для конкретных обследований, которые рассматриваются в этой главе. Кроме того, дается иллюстрация ряда других характеристик плана выборки, которые подробно рассматриваются в настоящем руководстве. К ним относятся, в том числе, неявная выборка, выбор оптимального размера кластера для снижения эффекта *deff* и распределение выборки по областям обследования.

4.2.6.2. Камбоджа, 1998–1999 годы

82. Эталонная выборка для Камбоджи демонстрирует использование двухэтапного плана выборки. Крупная выборка первичных единиц выборки предоставила эталонный список территориальных сегментов второго этапа, вошедших в подвыборки для использования в конкретных обследованиях.

83. Национальный институт статистики Камбоджи в 1999 году разработал эталонную выборку для использования в проводимой между переписями правительственной программе обследований домашних хозяйств, состоящей из периодического социально-экономического обследования и, возможно, обследований по вопросам здравоохранения, рабочей силы, доходов и расходов, демографии, а также внеплановых обследований. Перепись населения 1997 года послужила инструмен-

тарием для составления эталонной выборки, которое осуществлялось в два этапа. На первом этапе проводился отбор (с вероятностью, пропорциональной размеру) первичных единиц выборки, в качестве которых были определены деревни, при этом показателем размера являлась определенная переписью численность домохозяйств. Отбор первичных единиц выборки проводился с помощью компьютерной операции. Второй этап предусматривал создание территориальных сегментов в рамках отобранных первичных единиц выборки, и эта операция проводилась вручную.

84. Было принято решение в качестве первичных единиц выборки использовать деревни, поскольку они в среднем были достаточно крупными в плане численности домохозяйств (245 домохозяйств в городских районах и 155 — в сельских) для обслуживания нескольких обследований в период между переписями. Следовательно, удастся избежать излишней нагрузки, связанной с многократным опросом одних и тех же респондентов. Был рассмотрен, но отклонен альтернативный вариант использования счетных участков переписи, поскольку их размер в среднем равнялся половине размера деревень. В эталонную выборку не были включены особые группы населения, как проживающие временно или находящиеся в специальных учреждениях, так и проживающие в военных казармах.

85. В эталонную выборку было отобрано в общей сложности 600 первичных единиц выборки, поскольку считалось, что это число обеспечит достаточно равномерное распределение по территории страны для того, чтобы адекватно представить в выборке все провинции. При формировании выборки была использована неявная стратификация, которая осуществлялась путем сортировки списка деревень в географической последовательности — городские и сельские районы в разбивке по провинциям, районам и общинам. Таким образом, эталонная выборка была автоматически пропорционально распределена по городским и сельским районам и провинциям.

86. Интересной особенностью эталонной выборки Камбоджи стал второй этап формирования выборки. Как уже отмечалось, он предусматривал создание территориальных сегментов в рамках отобранных первичных единиц выборки. Необходимо подчеркнуть, что концепцию второго этапа формирования эталонной выборки не следует путать со вторым этапом формирования выборки, относящимся к отбору домохозяйств для конкретных обследований. В рамках каждой отобранной первичной единицы эталонной выборки были сформированы территориальные сегменты в размере (в среднем) 10 домохозяйств каждый, причем эта операция была канцелярской и проводилась в офисе. В этом контексте данная операция не предполагала никаких работ на местах, за исключением необычных случаев, поскольку были использованы имеющиеся списочные журналы и картосхемы переписи 1997 года. Число сегментов, создаваемых в каждой первичной единице эталонной выборки, рассчитывалось как число вошедших в перепись домохозяйств, деленное на 10 и округленное до ближайшего целого числа. Например, деревня, в которой по данным переписи 1997 года, насчитывалось, 187 домохозяйств, делилась на 19 сегментов.

87. Созданные таким способом сегменты стали структурными элементами с целью включения в выборку или подвыборку для применения в определенных обследованиях. Выбор одного или более сегментов из всего набора или из части первичных единиц выборки проводится для каждого обследования или раунда обследования, в котором используется эталонная выборка. Важным преимуществом здесь является то, что создание эталонной выборки описанным выше способом дает возможность для каждого использующего ее конкретного обследования стать самовзвешенным, в зависимости от характеристик его плана выборки.

88. Основным преимуществом плана эталонной выборки является то, что он обеспечивает бóльшую гибкость с точки зрения способов разбивки на подвыборки для использования в конкретных обследованиях. Отбор кластеров (иными словами — сегментов) для каждого обследо-

вания может, при желании, давать различный их набор. Типовая первичная единица выборки содержит около 18–30 сегментов, что обеспечивает вполне достаточное число сегментов в каждой первичной единице выборки для обслуживания всех обследований. Более того, становится возможным повторное проведение социально-экономического обследования каждый год с различным набором сегментов. В качестве альтернативы возможно также использовать метод частичного совпадения выборок путем сохранения каждый год определенной доли сегментов (скажем, 25 процентов) по методу ротации, причем ежегодно будут заменяться 75 процентов сегментов.

89. Один из недостатков плана эталонной выборки обусловлен необходимостью использования компактных кластеров (все домохозяйства в сегменте выборки расположены близко друг к другу). Это несколько увеличивает эффект схемы по сравнению с некомпактными сегментами, иными словами — с систематической выборкой домохозяйств в рамках более крупного кластера, а поскольку считается, что эффект схемы будет в определенной степени снижен за счет ограничения размера кластера, было решено использовать сегменты размером только в 10 домохозяйств.

90. Предполагалось, что обновление выборки будет производиться каждые два или три года. Хотя было признано, что лучше обновлять всю эталонную выборку, было решено проводить обновление только в пределах первичных единиц выборки, предназначенных для определенного запланированного на тот момент выборочного обследования. Обновление заключалось в выездах на местность, осуществляемых с целью подготовки новых списков домохозяйств в соответствующих сегментах. Повторный список составлялся для той же самой территории в пределах тех сегментов, которые содержали первоначальный набор домохозяйств, и это стало одной из причин особой важности установления границ сегмента.

91. Интересно отметить, что при проведении операции обновления в процессе создания камбоджийского плана эталонной выборки было налажено сотрудничество со старостами деревень. Было известно, что они тщательно ведут записи по всем домохозяйствам своих деревень, и, более того, эти записи регулярно обновляются. В большинстве случаев их записи оказывались достаточно точными. Кроме того, старосты деревень могли оказать неоценимую помощь в выявлении и нахождении территорий, подходящих для любого конкретного сегмента.

4.2.6.3. Объединенные Арабские Эмираты, 1999 год

92. Эталонная выборка Объединенных Арабских Эмиратов имеет две важных особенности своего формата. Во-первых, в плане эталонной выборки применена специальная стратификация для охвата двух различных групп населения — лиц, являющихся и не являющихся гражданами Объединенных Арабских Эмиратов. Во-вторых, план демонстрирует, каким образом можно использовать план стандартных сегментов (см. раздел 3.8.2) для решения проблем, связанных с различными размерами счетных участков и с использованием данных старой переписи.

93. Эталонная выборка Объединенных Арабских Эмиратов описана Министерством планирования как супервыборка из 500 первичных единиц выборки, основанная на переписи населения 1995 года в качестве инструментария выборки. Она предназначена для использования в конкретных обследованиях домашних хозяйств вплоть до момента проведения очередной переписи населения. В качестве первичных единиц выборки определены счетные участки переписи или их части, с тем чтобы каждая первичная единица выборки содержала в среднем около 60 домохозяйств, состоящих из лиц, как имеющих, так и не имеющих гражданство страны.

94. До начала отбора первичных единиц выборки были сформированы две страты. Первая состояла из счетных участков, в которых на момент переписи одна треть или более домохозяйств являлись домохозяйствами граждан страны. Вторая страта включала все прочие счетные участки.

В общей сложности 1 686 счетных участков было отнесено к страте I и 2 986 — к страте II. Используя систематический отбор с вероятностью, пропорциональной размеру, в каждой из двух страт была сформирована выборка из 250 первичных единиц выборки для общего числа 500 первичных единиц выборки по всей территории страны. Ожидалось, что эта эталонная выборка даст примерно одинаковое количество домохозяйств, относящихся к лицам, являющимся и не являющимся гражданами страны. Сначала была проведена сегментация крупных первичных единиц выборки (имеющих 90 и более домохозяйств), и в рамках каждой такой крупной первичной единицы выборки был произвольно выбран один сегмент. Для обновления инструментария в 500 первичных единицах выборки были составлены новые текущие списки домохозяйств. В результате операции по составлению списков в первичных единицах выборки было получено примерно 30 000 домохозяйств для использования в различных комбинациях для проведения конкретных обследований. Для обеспечения гибкости использования домохозяйства в каждой первичной единице выборки были разделены на 12 подгрупп или «наборов» примерно по 5 домохозяйств каждая.

95. Заслуживающей внимания особенностью этой эталонной выборки является то, что она не является самовзвешенной, поскольку две страты не равны по размеру. Первым обследованием, в котором использовалась эталонная выборка, стало Национальное обследование на предмет выявления диабета 1999 года, организованное Министерством здравоохранения. Ожидается проведение на местах и других обследований, включая обследование рабочей силы и обследование доходов и расходов (или семейного бюджета).

96. Ниже приводится еще ряд характеристик, связанных с определенными особенностями Объединенных Арабских Эмиратов, которые диктовали создание плана эталонной выборки. При этом особо тщательно учитывались два первостепенных фактора — целевые совокупности и инструментарий выборки.

97. Как отмечалось выше, в стране присутствуют две важных обследуемых группы населения — граждане и лица, не имеющие гражданства страны. Хотя, согласно переписи населения 1995 года, первые составляли около 43 процентов населения, на них приходилось всего лишь около одной четверти домохозяйств страны. Это было обусловлено тем, что домохозяйства лиц, не имеющих гражданства страны, были значительно меньше с точки зрения численности лиц, проживающих в одном домохозяйстве. Последствием такого факта для плана выборки стало то, что при пропорциональном формировании выборки домохозяйств страны, почти три четверти участвующих в опросе домохозяйств относились бы к лицам, не являющимся гражданами. Еще одним последствием стало то, что надежность итоговых оценок обследования для домохозяйств лиц-неграждан была бы примерно в три раза выше, чем для домохозяйств граждан страны. Поскольку результаты обследования предназначались для разработки политики и планирования программ, такое расхождение в надежности оценок было сочтено нежелательным и непригодным. Подходящим решением с точки зрения плана выборки стала обработка этих двух несопоставимых и неравных обследуемых совокупностей как отдельных групп с помощью применения надлежащей, описанной выше стратификации.

98. Был использован также и другой уровень стратификации, а именно территориальная стратификация для обеспечения надлежащего распределения выборки по отдельным эмиратам и по городским и сельским районам. До формирования выборки список счетных участков был отсортирован в следующей последовательности: сначала страты граждан в разбивке на городское население, а в рамках городского населения в разбивке по эмиратам и в рамках каждого эмирата в разбивке на коды счетных участков, расставленных в порядке убывания доли граждан, далее в разбивке на сельское население, по эмиратам и по кодам счетных участков; затем страты неграждан страны в той же последовательности.

99. Было признано, что основной характеристикой инструментария эталонной выборки должен быть набор карт, на которых обозначены границы территориальных единиц, определенных в качестве районов выборки или, иными словами, первичных единиц выборки. Территориальные единицы должны быть достаточно малыми, чтобы удобнее было составлять списки, но в то же время достаточно большими, чтобы их можно было четко привязать к естественным границам (для легкого определения места нахождения). Счетные участки переписи были признаны единственными приемлемыми с практической точки зрения территориальными единицами, которые отвечали этим двойным критериям. К сожалению, в ходе переписи населения карты не использовались, и поэтому существующие счетные участки не были привязаны к известным границам. В результате возникла необходимость обеспечить составление достоверной информации о границах первичных единиц эталонной выборки (счетных участков).

100. При подготовке эталонной выборки из первичных единиц выборки использовался «план стандартных сегментов», описанный в главе 3. Эта методология успешно применялась во многих странах, как в Программе обследований в области народонаселения и здравоохранения, так и Панарабского обследования материнского и детского здоровья (ПАПЧАЙЛД).

101. Было решено использовать план стандартных сегментов, поскольку счетные участки переписи в Объединенных Арабских Эмиратах достаточно сильно варьируются по размерам. Были созданы стандартные сегменты примерно по 60 домохозяйств. Число стандартных сегментов в каждой первичной единице выборки рассчитывалось как общее число домохозяйств (как граждан, так и неграждан), деленное на 60 и округленное до ближайшего целого числа.

102. В тех случаях, когда число сегментов, иными словами, показатель размера составлял два и более, первичная единица выборки делилась на территориальные сегменты. Это требовало проведения операции на местах, которая предусматривала посещение счетного участка (первичной единицы выборки) и подготовки картосхемы с помощью быстрого подсчета и нанесения на карту жилых помещений (не домохозяйств). После сегментации из каждой первичной единицы выборки произвольно отбирался один сегмент. Этот сегмент становился фактическим территориальным районом для формирования эталонной выборки. Еще один выезд на место осуществлялся для составления текущего списка домохозяйств по каждому сегменту выборки. Таким образом, последняя процедура, необходимая для обновления инструментария эталонной выработки трехлетней давности, рассматривалась как основной элемент операции по формированию выборки.

103. Итоговая операция формирования эталонной выборки заключалась в разделении включенных в обновленный список домохозяйств в каждом сегменте выборки на 12 систематических подгрупп или «наборов». Для конкретных обследований должен использоваться один или несколько наборов. Поскольку средний размер сегмента равнялся примерно 60 домохозяйствам, каждый набор содержал в среднем по 5 домохозяйств.

104. Основной причиной составления именно 12 наборов стала та гибкость, которая обеспечивалась этим числом наборов при формировании различных комбинаций для использования в обследованиях. Фактический выбор для того или иного обследования зависел от различных факторов, включая цели обследования, желаемый размер кластера и общий размер выборки, который требовался для обследования. Например, две пятых первичных единиц выборки были использованы в Национальном обследовании по проблеме диабета. Следовательно, в него включались 4 из 12 наборов домохозяйств в рамках этих первичных единиц выборки. Эта комбинация позволила создать общий план выборки из 200 первичных единиц с кластерами размера 20 (т. е. 4 раза по 5 домохозяйств) и общим размером выборки приблизительно в 4 000 домохозяйств.

4.2.6.4. Вьетнам, 2001 год

105. Эталонная выборка Вьетнама имеет две отличительные особенности. Она демонстрирует использование двух этапов при отборе эталонной выборки и третьего этапа в случае применения для конкретных обследований. Во-вторых, она демонстрирует, каким образом эталонная выборка может быть распределена между географическими областями.

106. Эталонная выборка, которая базируется на переписи 1999 года в качестве инструментария выборки, включает двухэтапный план. В качестве первичных единиц выборки были определены общины в сельской местности и микрорайоны в городах. Такое определение было обусловлено принятым решением о том, что в каждой первичной единице выборки необходимо наличие не менее 300 домохозяйств, чтобы выступать в качестве эталонной выборки. В качестве альтернативного варианта рассматривалось использование счетных участков как первичных единиц выборки, однако они оказались слишком малыми по размеру, и возникала необходимость объединения ряда соседних счетных участков для надлежащего соответствия критериям первичных единиц выборки. Эта задача была признана крайне громоздкой и трудоемкой. С другой стороны, число общин/микрорайонов, которые было необходимо объединить из-за их малого размера, составляло всего 529 из более чем 10 000.

107. Всего для эталонной выборки было отобрано 3 000 первичных единиц выборки с помощью использования вероятности, пропорциональной размеру. Каждая первичная единица выборки содержала в среднем по 25 счетных участков в городах и по 14 — в сельских районах. Для второго этапа отбора в каждой первичной единице выборки методом вероятности, пропорциональной размеру, были отобраны три счетных участка. Эти единицы второго этапа — т. е. счетные участки — содержали в среднем примерно по 100 домохозяйств в соответствии с данными переписи 1999 года: по 105 — в городах и по 99 — в сельских районах.

108. Целью эталонной выборки было предоставление надежных данных по каждому из восьми географических регионов Вьетнама. Формирование выборки осуществлялось независимо в каждой провинции. Таким образом, провинции служили в качестве страт для эталонной выборки. В некоторых малых провинциях с очень небольшой численностью населения было признано желательным провести дополнительную выборку. Соответственно, распределение выборки между провинциями осуществлялось методом вероятности, пропорциональной квадратному корню размера провинции. Для городских и сельских районов было использовано пропорциональное распределение.

109. В дополнение к указанной выше стратификации на уровне провинций, осуществлялась неявная территориальная стратификация в пределах провинций. Для использования эталонной выборки в конкретных обследованиях предполагалось применение подгрупп счетных участков: например, одна треть из них — для Многоцелевого обследования домашних хозяйств. Для применения в обследованиях организуется третий этап отбора, при котором из каждого вошедшего в выборку счетного участка отбирается фиксированное число домохозяйств. Это число может варьироваться для разных обследований и для городских и сельских районов. Например, могут отбираться по 20 домохозяйств в каждом сельском *счетном участке* и по 10 — в каждом городском *счетном участке*.

4.2.6.5. Мозамбик, 1998–1999 годы

110. Эталонная выборка Мозамбика иллюстрирует тот случай, когда предполагалось использование одноэтапного формирования первичных единиц выборки для всех организуемых прави-

тельством национальных обследований в рамках программы обследований домашних хозяйств страны в период между переписями. Она также иллюстрирует, каким образом гибкая эталонная выборка может адаптироваться для соответствия целям измерения показателей конкретного обследования.

111. Первичные единицы эталонной выборки в Мозамбике были определены и сформированы наиболее простым путем, как описано в различных разделах настоящего руководства. Первичные единицы выборки формировались на базе переписи населения 1997 года, которая использовалась в качестве инструментария выборки. Они состояли из территориальных групп размером, как правило, от трех до семи счетных участков, которые включали в среднем примерно по 100 домохозяйств. Первичные единицы эталонной выборки отбирались с использованием вероятности, пропорциональной размеру.

112. Всего было отобрано 1 511 первичных единиц выборки для создания структуры выборки, которую предполагалось применять в единой системе обследований домашних хозяйств Мозамбика. Первичные единицы эталонной выборки были разделены на группы, каждая из которых представляла собой систематическое подмножество, а следовательно, сама по себе являлась вероятностной выборкой. Существует 10 таких групп примерно по 151 первичной единице выборки в каждой. В рамках пятилетнего плана (2000–2004 годы) первым обследованием, в котором используется данная эталонная выборка, стал разработанный Всемирным банком «Вопросник по основным показателям благосостояния».

113. План выборки для Вопросника по основным показателям благосостояния был разработан с учетом двух целей проведения измерений. Первая — получение актуальных показателей, необходимых для составления профиля благосостояния людей и домохозяйств в Мозамбике. Вторая — предоставление надежных оценок этих показателей на национальном уровне, отдельно для городских и сельских районов, а также для каждой из 11 провинций страны. В методике формирования выборки для Вопросника по основным показателям благосостояния использовалась эталонная выборка Мозамбика, в рамках которой отбиралось примерно 14 500 домохозяйств в стратифицированном, объединенном в кластеры формате. Соответственно, первым этапом отбора стали, естественно, первичные единицы выборки.

114. Вторым этапом отбора для Вопросника по основным показателям благосостояния стала подвыборка первичных единиц эталонной выборки. Всего в эту подвыборку вошло 675 первичных единиц выборки из содержащихся в ней 1 511 единиц. Они отбирались систематически по принципу равной вероятности и, в дополнение к этому, были равномерно распределены по 11 провинциям Мозамбика. На третьем этапе формировалась выборка одного счетного участка в каждой из первичных единиц выборки, включенной в Вопросник по основным показателям благосостояния. Таким образом, в наличии имелось 675 кластеров в составе выборки Вопросника — 475 сельских и 200 городских. Счетные участки отбирались с равной вероятностью, поскольку имели примерно равный размер — как указывалось выше, в среднем примерно 100 домохозяйств, хотя и присутствовала определенная вариантность. Конечный этап отбора проводился по итогам работ на местах, в рамках которых сотрудники Национального статистического института [Instituto Nacional de Estatistica (INE)] посещали кластеры для составления новых списков домохозяйств в целях обновления инструментария выборки 1997 года. Из составленных таким образом списков была сформирована систематическая выборка из 20 домохозяйств в сельских районах и 25 — в городских для проведения обзорных опросов по Вопроснику. Следовательно, формирование выборки для Вопросника представляло собой четырехэтапный процесс отбора, несмотря на то что эталонная выборка, на которой этот процесс базировался, была одноэтапной.

115. Две особенности Вопросника по основным показателям благосостояния иллюстрируют гибкость, с которой эта эталонная выборка может адаптироваться к конкретным требованиям применения в том или ином обследовании.

116. Во-первых, при применении эталонной выборки для Вопросника по основным показателям благосостояния Национальный институт статистики был заинтересован использовать те группы, которые, как указывалось выше, уже были определены. Учитывая желание иметь около 600 первичных единиц выборки для Вопросника, ожидалось, что можно будет использовать четыре группы. Эта идея, однако, была отклонена, когда выяснилось, что число необходимых для Вопросника групп будет разным для каждой провинции, поскольку цели измерения показателей требовали примерно одинакового размера выборки по провинциям. Вместо этого необходимые для Вопросника 675 первичных единиц выборки были систематически отобраны из всего файла эталонной выборки независимо от групп. Такой отход от первоначальных целей эталонной выборки, а именно — предоставить пропорциональную выборку по провинциям, был обусловлен желанием получить для Вопросника равную по провинциям надежность оценок в отличие от первоначального плана эталонной выборки. В первоначальном плане эталонной выборки превалирующее значение, как ожидалось, должны были иметь оценки на национальном уровне.

117. Второй важной проблемой плана выборки для Вопросника стал размер кластеров. Было согласовано, что размеры кластеров будут разными для городских и сельских домохозяйств на том основании, что эффект схемы или *deff* был выше в сельских районах, где средства к существованию с экономической точки зрения обеспечивались за счет натурального сельского хозяйства. Иными словами, домохозяйства в сельских районах с большой вероятностью должны иметь аналогичные характеристики. Эталонная выборка на конечном этапе давала возможность отбора различного фиксированного числа домохозяйств (25 и 20 для городских и сельских районов, соответственно).

4.3. Краткие рекомендации

118. В данном разделе приводится резюме основных рекомендаций, которые следует выделить из этой главы. Они представлены в формате контрольного списка и, как в главе 3, должны восприниматься скорее как «эмпирические правила», нежели как непреложные рекомендации. К ним относятся следующие:

- использовать инструментарии выборки, которые являются по возможности полными, точными и обновленными;
- обеспечивать, чтобы инструментарий выборки охватывал планируемую обследуемую совокупность;
- по мере возможности, использовать в качестве инструментария для обследований домашних хозяйств результаты последней переписи;
- определять первичные единицы выборки в инструментарии в виде территориальных единиц, таких как *счетные участки* переписи с нанесенными на карту и четко определенными границами, для которых имеются данные по численности населения;
- использовать список домохозяйств по данным переписи в качестве инструментария на последнем этапе, только если этот список совсем недавно обновлялся — обычно не более года назад;

- использовать двойные или множественные инструментарии с осторожностью, обеспечивая при этом функционирование процедур контроля за дублированием данных;
- обновлять инструментарий переписи, если ему больше двух лет, на национальном уровне или по конкретным целевым районам, известным своими быстрыми темпами роста с помощью:
- обхода территории с быстрым подсчетом для обновления старого инструментария;
- обновления кластеров выборки путем составления новых списков домохозяйств;
- обновлять данные только по кластерам выборки, если инструментарию переписи не более двух лет путем:
- составления обновленных списков домохозяйств;
- использовать эталонную выборку или инструментарий эталонной выборки только тогда, когда планируется или осуществляется крупномасштабная программа непрерывных обследований;
- формировать в качестве первичных единиц эталонной выборки только единицы, достаточно крупные или многочисленные для обслуживания многих обследований или повторных раундов обследований в течение периода между переписями;
- обновлять инструментарий эталонной выборки, используя те же рекомендации, которые предложены выше для инструментариев одноразовых обследований;
- применять систему ротации выборки — либо домохозяйств, либо *первичных единиц выборки* — в повторных обследованиях, использующих эталонные выборки.

Справочная литература и дополнительные источники информации

Cochran, W. G. (1977), *Sampling Techniques*, 3rd ed., New York: Wiley.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow (1953). *Sample Survey Methods and Theory*, New York: Wiley.

International Statistical Institute (1975). *Manual on Sample Design*. World Fertility Survey Basic Documentation, Voorburg, Netherlands.

Kalton, G. (1983). *Introduction to Survey Sampling*, Beverly Hills, California: Sage. Publications Kish, L. (1965), *Survey Sampling*. New York: Wiley.

League of Arab States (1990). *Sampling Manual, Arab Maternal and Child Health Survey*, Basic Documentation 5. Cairo: Pan Arab Project for Child Development (PAPCHILD).

Macro International, Inc. (1996). *Sampling Manual*. DHS-III Basic Documentation, No. 6. Calverton, Maryland.

Pettersson, Hans. (2001), Mission report: recommendations regarding design of master sample for household surveys of Viet Nam. Unpublished. General Statistical Office, Hanoi. 25 November.

_____ (2005) «Планирование инструментариев эталонной выборки и эталонных выборок для обследований домашних хозяйств в развивающихся странах» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96, в продаже под № R.05.XVII.6.

Turner, A. (1998), Mission report to the Kingdom of Cambodia, National Institute of Statistics, 11–24 November. National Institute of Statistics, Phnom Penh: Unpublished.

_____ (1999), Mission report to United Arab Emirates, Ministry of Health and Central Department of Statistics 23 January–3 February. Abu Dhabi: Unpublished. Central Department of Statistics.

_____ (2000). Mission report to Mozambique, Instituto Nacional de Estatistica, 13–26 August. Instituto Nacional de Estatistica, Maputo: Unpublished.

Статистический отдел Организации Объединенных Наций (1984). *Руководство по обследованию домашних хозяйств* (Переработанное издание). Методологические обследования, № 31; в продаже под № R.83.XVII.13.

_____ (1986), National household Survey Capability Programme: *Sampling frames and sample designs for integrated household survey programmes*. Preliminary Version. DP/UN/INT-84-014/5E. New York: United Nations Department of Technical Cooperation for Development and Statistical Office.

United Nations Children's Fund (2000). *End-Decade Multiple Indicator Survey Manual*. New York: UNICEF chapter. 4.

United States Bureau of the Census (1978). *Current Population Survey Design and Methodology*. Technical Paper 40. Washington, D.C.: Bureau of the Census.

Verma, Vijay (1991). *Sampling Methods*. Training Handbook. Tokyo: Statistical Institute for Asia and the Pacific.

World Bank (1999). *Core Welfare Indicators Questionnaire (CWIQ) Handbook*. Washington D.C.: World Bank. chapter. 4.

Глава 5

Документирование и оценка планов выборки

5.1. Введение

1. Настоящая глава, хоть и сравнительно краткая, играет в этом пособии центральную роль. Как правило, на фоне аврального режима публикации результатов обследования вопросы документирования и оценки планов выборки, в частности, и методологии обследования, в общем, слишком часто упускаются из виду. Это особенно справедливо для стран, имеющих лишь небольшой предшествующий опыт проведения обследований домашних хозяйств, и проявляется в том, что метаданные подчас бывают ненадлежащим образом представлены в рабочих таблицах и отчетах обследования. В некоторых странах недооценивается значение документирования процедур обследования, включая осуществление выборки. В связи с этим, там отсутствуют традиции, требующие такой работы. Несмотря на то что исследователи проводят анализ полученных по результатам обследования данных, они, тем не менее, должны быть заинтересованы в ознакомлении с методологиями, связанными с составлением плана выборки. Следовательно, возникает необходимость в подготовке целого ряда документов с подробным изложением соответствующих процедур, используемых в обследовании.

2. Оценка результатов обследования зачастую вообще игнорируется. Вследствие этого в анализ обследования закрадываются ошибки. Обычно такое положение дел обусловлено тем фактом, что бюджетные ограничения часто не позволяют провести какие-либо формальные исследования или разработать методологии оценки многочисленных ошибок регистрации данных, которые могут появиться в обследованиях домашних хозяйств. При этом существуют и другие «барометры» качества данных (такие, как доля неполучения ответов), которые должны быть легкодоступными, но и они зачастую не упоминаются в отчетах обследований.

3. В настоящей главе также подчеркивается важность предоставления пользователям соответствующей информации об известных недостатках полученных данных, даже когда формальные исследования оценок не проводились. В этой связи важно отметить, что в цели данной главы не входит обсуждение многочисленных принятых технологий проведения формальных оценок методологии обследования[1]. Вместо этого, основное внимание в данной главе уделено тому, какую информацию необходимо предоставлять пользователям для оказания им помощи в оценке качества обследования с упором на различные аспекты выборки.

1 К специальным исследованиям, предназначенным для оценки конкретных видов ошибок регистрации данных в обследованиях, относятся повторные опросы (для оценки вариантности ответов), обследования после проведения переписи (для оценки охвата и содержания), взаимопроникающие выборки (для оценки фактора вариантности, обусловленного действиями регистратора), проверка записей в обратном порядке (для оценки ошибок вспоминания респондентов) и т. д.

5.2. Необходимость в проведении и виды документирования и оценок выборки

4. Для обследований домашних хозяйств необходимы два вида документации. Первый вид состоит из подробной отчетности об обследовании и оперативном проведении процедур формирования выборки в процессе обследования. Без такой документации в анализ обследования закрадываются ошибки. Например, без тщательного ведения отчетности на момент проведения анализа могут не быть в полной мере известны значения вероятности отбора.

5. Именно поэтому технический специалист по выборке должен принять необходимые меры к тщательному документированию не только плана выборки для проведения конкретного обследования, но и осуществления этого плана. Планы выборки зачастую требуют адаптации на различных этапах работ на местах по причине непредвиденных ситуаций, возникающих при проведении обследования. Важно документально фиксировать шаг за шагом все процедуры, используемые при осуществлении плана выборки, с тем чтобы удостовериться в том, что они проводятся строго по этому плану. Если этого не происходит, то еще более важно фиксировать все отклонения от плана, даже самые незначительные. Эта информация пригодится позднее, на этапе анализа, если потребуется вносить какие-либо корректировки. Более того, такая документация незаменима при планировании будущих обследований.

6. Второй вид документации представляет собой отчеты. Для каждого обследования домашних хозяйств необходимо готовить два типа технических отчетов: сравнительно краткое, удобное для пользователя описание методологии обследования, включая план выборки и его осуществление; а также более детальное описание методологии обследования. Первое из указанных описаний, как правило, включает «технические» разделы (или добавления) из различных предметных отчетов, опубликованных для обсуждения и интерпретации выводов по существу предмета обследования[2], включая подраздел о том, что известно об ограничениях полученных данных (см. ниже).

7. Второй вид технических отчетов предназначен в большей степени для профессиональных исследователей, социологов и статистиков, чем для лиц, принимающих решения, или широкой общественности, причем они должны содержать более подробное описание методологии и обследования и должны скорее стоять особняком, нежели входить в качестве составной части в серию отчетов по существу вопросов. Бюро переписей Соединенных Штатов (1978 год) опубликовало отличный вариант такого отчета. Конечно, предпочтительно иметь одновременно и детальный технический отчет, и обычные отчеты о переписи, хотя первый, как правило, готовится значительно позже, а иногда и вообще опускается. Целесообразно также опубликовать технический отчет или его сокращенный вариант в каком-либо статистическом журнале для обеспечения его долговечности.

8. Отчеты обоих видов настолько важны, что национальным статистическим бюро рекомендуют поручить специальному отделу или сотруднику на регулярной основе готовить их для различных обследований домашних хозяйств.

2 Рекомендации о том, что необходимо включать в доклад о выводах из выборочного обследования, содержатся в докладе Подкомиссии по статистической выборке Организации Объединенных Наций (Организация Объединенных Наций, 1964 год).

5.3. Маркировки для переменных параметров плана выборки

9. В данном разделе, а также в разделах 5.2–5.7 рассматривается документация первого вида, которая предусматривает ведение отчетности по связанным с выборкой процессам обследования.

10. Определенные для каждого этапа единицы отбора должны иметь ясную и уникальную маркировку. При многоэтапной планировке это будет означать введение кодов для первичных, вторичных, третичных и конечных единиц выборки (в зависимости от числа этапов в плане). Как правило, четырехзначного кода будет достаточно для первого этапа отбора и трехзначного кода — для остальных этапов. Географические области также следует надлежащим образом маркировать. В дополнение к этому, в процессе маркировки должны участвовать административные коды, обозначающие административно-территориальную структуру тех районов, к которым относятся единицы выборки. Также должны четко маркироваться аналитические единицы.

Пример

Предположим, что в соответствии с двухэтапным планом формируется выборка из 1 200 первичных единиц, в качестве которых выбраны счетные участки, по 600 единиц в каждой из двух определенных областей обследования — городских и сельских. Удобным способом кодирования первичных единиц выборки будет нумерация кодов с 0001 до 1 200. Более того, целесообразно присваивать эти коды в той же последовательности, которая использовалась для отбора первичных единиц выборки. Эта особенность может понадобиться для использовании при расчете величин вариантности выборки. Таким образом, если первыми были отобраны первичные единицы выборки в сельских районах, они получат коды от 0001 до 0600, а городские единицы будут обозначаться кодами от 0601 до 1 200. Такая система кодирования имеет два преимущества: во-первых, каждая первичная единица выборки получает уникальный номер и маркировку, а, во-вторых, аналитики смогут с первого взгляда определить, является ли та или иная первичная единица выборки городской или сельской, просто исходя из ее идентификационного кода. На втором этапе формирования выборки для каждой первичной единицы выборки составляется список, из которого 20 домохозяйств отбираются для опроса. На этом этапе всем вошедшим в список домохозяйствам присваивается трехзначный код (или четырехзначный — если некоторые счетные участки содержат свыше 999 домохозяйств), снова в той последовательности, в которой они обозначены в списке. Поскольку вошедшие в выборку домохозяйства сохраняют присвоенные им таким образом коды, то отобранным для опроса домохозяйствам новые коды (например, от 1 до 20) даваться не будут. И наконец, административные коды присваиваются по мере необходимости. Следовательно, код 09 003 008 0128 для вошедшего в выборку домохозяйства 080 будет обозначать, что это домохозяйство имеет восьмидесятый номер по списку (и отобрано для опроса) и входит в первичную единицу выборки 0128, которая относится к гражданскому административному подразделению 008 в районе 003 провинции 09. Более того, номер первичной единицы выборки сразу же говорит о том, что оно относится к сельской области. Если в ходе обследования собирается информация о членах домохозяйства, то каждый из его членов также получит уникальный двухзначный код от 01 до 99.

11. Надлежащая маркировка необходима, во-первых, для контроля качества: по мере того, как регистраторам даются задания, и вопросники возвращаются с мест опроса, они могут сверяться с эталонным списком, чтобы удостовериться в охвате всех вошедших в выборку домохозяйств. Во-вторых, системы уникальных номеров неоценимы для сотрудников по обработке данных, поскольку они позволяют составлять таблицы в соответствии с географическим местоположением.

12. Для стран, осуществляющих программы многих обследований, весьма желательно, чтобы все переменные параметры плана выборки маркировались в последовательном, стандартизованном порядке во всех обследованиях. Тем самым устраняется путаница в данных, как у производителей, так и у пользователей, и это дает очевидные преимущества в процессе обработки данных и представления результатов.

13. Что касается последнего пункта, то состоящая из многих обследований программа только выиграет от присвоения кодов всей генеральной совокупности первичных единиц выборки, как описано в предыдущем примере, а не только непосредственно вошедшим в выборку единицам. Это обусловлено тем, что различные первичные единицы часто входят в выборку различных обследований, причем такая ситуация в особенности относится к случаям применения эталонной выборки.

14. Как правило, эталонные выборки нуждаются в маркировке плана даже больше, чем выборки одноразовых обследований. Как обсуждалось ранее, ключевой сферой применения эталонных выборок являются повторные раунды одного и того же обследования. Надлежащая маркировка переменных параметров плана выборки, обозначающая этапы отбора, абсолютно необходима для отслеживания случаев частичного перекрытия выборки одного обследования со следующим. Часто вводятся ротационные группы (систематические подсовокупности полной выборки) в целях облегчения идентификации единиц (домохозяйств, кластеров или ПЕВ), которые меняются в последующих раундах обследования. Такие группы, безусловно, требуют собственных идентификационных кодов. Более того, должны получать соответствующие коды и те домохозяйства, которые добавляются к эталонной выборке в ходе ее периодического обновления. Схема кодирования должна быть разработана таким образом, чтобы различать новые и старые домохозяйства.

5.4. Вероятности отбора

15. Одним из информационных пунктов, который зачастую упускается в документации выборки, является запись вероятностей отбора на различных его этапах. Если такая информация и присутствует, то часто она ограничена общим весовым коэффициентом выборки (из которого можно быстро рассчитать полную вероятность) для каждой конкретной выборки.

16. Особенно важно подробно отражать этот аспект в документации при формировании подвыборок на местах в ходе работ по сбору данных, что может произойти в случае чрезмерно крупного сегмента/кластера выборки. Например, как рассматривалось выше, может возникнуть необходимость в разбивке на сегменты непредвиденно крупного кластера, скажем, на четыре части примерно равного размера. Затем одна из частей в случайном порядке отбирается для включения в список и проведения опроса. В этом случае полная вероятность вошедшего в выборку сегмента (а также отобранных в его рамках домохозяйств/лиц) равняется одной четвертой от первоначального кластера; а его вес, таким образом, является обратной величиной от одной четвертой, т. е. равняется четырем. Такой весовой коэффициент должен отражаться в расчетах в ходе анализа данных.

17. Формирование подвыборки может также иметь место, если в одном жилом помещении находятся более одного домохозяйства (когда жилое помещение является единицей списка). Одним из вариантов, который не дает смещения, является решение руководителя обследования о проведении опроса всех обнаруженных домохозяйств: так часто поступают, если имеются всего лишь два домохозяйства. Однако, если обнаруживается, скажем, пять домохозяйств там, где ожидалось одно, то соображения ограничения расходов могут продиктовать решение, что должен проводиться опрос только одного из них, естественно, выбранного в случайном порядке. И вновь чрезвычайно важна тщательная регистрация доли подвыборки (в этом примере 1/5), с тем чтобы

ответственные за выборку сотрудники могли точно рассчитать вероятность отбора данного домохозяйства и надлежащим образом скорректировать (с помощью коэффициента 5) его вес.

18. Целесообразно также документально фиксировать вероятности отбора на каждом его этапе, как отмечено в первом пункте данного раздела. Например, вероятность отбора каждой первичной единицы выборки будет иметь различную величину в любом случае применения формирования выборки с вероятностью, пропорциональной размеру. Это справедливо даже тогда, когда общий план выборки является самовзвешенным. Если вероятности отбора первичных единиц выборки игнорируются или документируются с ошибкой, может быть утрачена возможность определения общих весовых коэффициентов выборки. Для точного определения процедур подвыборки полезно знать, например, какими были первоначальные вероятности отбора.

5.5. Значения доли ответивших на вопросы и уровней охвата на различных этапах формирования выборки

19. В качестве составной части процесса оценки проведения выборочного обследования чрезвычайно важно предоставлять пользователям информацию о доле ответивших на вопросы и уровне охвата. Такую информацию целесообразно предоставлять в максимально подробном виде. Следовательно, важно давать не только сведения о доле ответивших на вопросы (или дополняющий его показатель — долю неполучения ответов), но также и таблицу причин неполучения ответов. В категории неполучения ответов, вероятнее всего, войдут следующие причины:

- отсутствие жильцов дома
- незаселенная жилая единица
- снесенная или непригодная для проживания жилая единица
- отказ от ответа
- временное отсутствие (отпуск и т. д.).

20. Определение доли ответивших на вопросы в плане разбивки по категориям может варьироваться для разных стран. В типичном случае, однако, законченный ответ включает в себя категории один, четыре и пять из указанных выше. Эти категории включают случаи, когда ответ должен быть получен, если это вообще возможно. Незаселенные и снесенные жилищные единицы обычно игнорируются (при расчете доли ответивших) на том основании, что в таких единицах невозможно получить ответ по определению. Таким образом, в той или иной стране может пройти, например, отбор 5 000 домохозяйств со следующими результатами: 4 772 завершенных опроса, 75 случаев «отсутствия жильцов дома», 31 незаселенное жилое помещение, 17 снесенных жилищных единиц, 12 отказов и 93 случая «временного отсутствия». Доля ответивших на вопросы будет рассчитываться, как обычно, исключая незаселенные и снесенные помещения, с результатом 4 772/(5 000–31–17) или 96,4 процента.

21. Когда целевые совокупности обследования включают как домохозяйства — для получения таких переменных параметров, как доход домохозяйства или доступ к услугам, так и людей (для изучения, например, состояния здоровья взрослого женского населения), принято рассчитывать доли ответивших на вопросы на уровне как домохозяйств, так и отдельных лиц. Например, ответы на вопросы могут представить 98 процентов домохозяйств, однако небольшая доля отдельных лиц в рамках этих домохозяйств может попасть в категорию нереспондентов.

22. Зачастую не проводится опрос целых кластеров по различным причинам, включая проблемы безопасности, например гражданские волнения или беспорядки, и отсутствие доступа из-за пересеченной местности или погодных условий. При возникновении таких проблем часто отбираются замещающие кластеры, причем такая процедура приводит к серьезным искажениям в оценках, поскольку население замещающих кластеров практически всегда значительно отличается по своим характеристикам от населения замещаемых кластеров. Тем не менее, когда такая замена производится, группа по проведению обследования обязана документально зафиксировать число и местоположение таких кластеров. Более того, в таких случаях важно также представить определенную информацию о неполном охвате. Это можно сделать путем проведения, по мере возможности, оценки численности людей в обследуемой группе(ах) населения, предположительно проживающих в конкретных районах, с разбивкой по представленным в выборке замещаемым кластерам.

23. Следует отметить, что проблемы типа упомянутых в предыдущем пункте можно сгладить в некоторой степени, если еще до формирования выборки выявить районы страны, находящиеся «за рамками охвата» опросов в ходе обследования в связи с проблемами безопасности или доступности. Такие выявленные районы необходимо документально оформить и исключить из генеральной совокупности обследования до формирования выборки. Исключенные районы должны быть отмечены в отчете вместе с сообщением о том, что результаты обследования неприменимы к исключенным районам.

5.6. Взвешивание: базовые весовые коэффициенты, корректировки на неполучение ответов и прочие корректировки

24. Способ расчета весов обследования приводится в главе 6. В настоящей главе подчеркивается важность документирования таких расчетов.

25. Взвешивание данных в обследованиях домашних хозяйств включает три операции: расчет базисных весов или весов схемы, внесение поправок на неполучение ответов и постстратификационных поправок. На практике во многих случаях используются только веса схемы, однако в некоторых случаях веса схемы могут корректироваться за счет учета дополнительного фактора, отражающего неполучение ответов. В сравнительно немногих случаях в ходе взвешивания может учитываться какой-либо другой фактор, как с поправкой на неполучение ответов, так и без нее, причем такой фактор предназначен для коррекции распределения совокупности, полученной за счет выборки, для того чтобы согласовать его с показателями распределения из другого независимого источника данных, например последней переписи. Такая процедура часто называется постстратификационным взвешиванием. В некоторых случаях взвешивание не производится вообще, это может произойти только при соблюдении двух условий: выборка должна быть полностью самовзвешенной, и получаемые данные должны ограничиваться процентными долями распределений в виде соотношений и коэффициентов, в отличие от оцененных суммарных или абсолютных величин.

26. При использовании взвешивания, конечно же, необходимо тщательно документировать все расчеты. Как указывалось выше, необходимо рассчитывать и документировать веса (или вероятности) на каждом этапе отбора. Кроме того, необходимо фиксировать веса отдельных показателей на всех этапах обработки данных, в частности: *a*) веса схемы, *b*) веса схемы после их умножения на поправочный коэффициент(ы), учитывающий неполучение ответов и *c*) веса схемы из пункта *b* после применения поправочных коэффициентов на постстратификацию.

27. Важно отметить, что веса схемы будут отличаться для каждой области обследования в любом случае, когда план выборки предусматривает проведение оценок по таким областям. Иными сло-

вами, даже при самовзвешенной выборке в пределах областей обследования каждая область будет иметь свой собственный отдельный вес. Более того, каждая область обследования будет иметь различную совокупность весов в том случае, если в пределах областей план выборки не является самовзвешенным. Необходимо также отметить, что поправки на неполучение ответов часто применяются отдельно по важным территориальным подразделениям, таким как крупные регионы, вне зависимости от того, присутствуют ли в плане выборки оценки по областям. И наконец, сам по себе вес схемы может умножаться на дополнительный коэффициент для определенных кластеров или домохозяйств. Это может осуществляться при любом применении подвыборки (см. раздел 5.3).

5.7. Информация по расходам на формирование выборки и проведение обследования

28. Несмотря на то что бюджет обследований домашних хозяйств обычно составляется весьма тщательно, важно, тем не менее, вести отчетность по фактическим расходам на различные операции в рамках обследований. Документирование расходов на обследование полезно, например, при составлении плана эталонных выборок и ведении отчетности по расходам на различные мероприятия в рамках обследования, а также важно при планировании будущих обследований.

29. При использовании эталонных выборок имеют место значительные первоначальные расходы, связанные с ее формированием. Такие расходы обычно включают следующие аспекты: *a*) компьютерная обработка файлов переписи для создания инструментария выборки, *b*) нанесение на карту и другие картографические работы по созданию первичных единиц выборки и *c*) отбор с помощью компьютера первичных единиц выборки. Как отмечалось в предыдущей главе, часть расходов на первоначальные операции зачастую берут на себя те министерства, которые будут использовать эталонную выборку в течение всего ее жизненного цикла. Эти расходы также необходимо распределять между всеми заранее известными обследованиями, для которых предназначена данная эталонная выборка. Таким образом, чрезвычайно важно тщательно вести отчетность по созданию эталонной выборки, а также по планированию всех аспектов выборки для будущих обследований.

30. Когда эталонная выборка готова, необходимо составить смету расходов на ее обновление. Как отмечалось ранее, обновление эталонной выборки проводится через определенные периоды времени, и соответствующие расходы, конечно, должны тщательно контролироваться.

31. На регулярной основе требуется получать данные по статьям расходов на формирование выборки, которые указаны в приводимом ниже списке, причем этот список применим в отношении как выборки для одноразовых обследований, так и эталонной выборки:

a) заработная плата сотрудников, занятых формированием выборки, включая гонорары всех внешних консультантов;

b) расходы по работам на местах в целях обновления инструментария выборки, включая расходы на персонал и подготовку вспомогательных материалов, таких как карты;

c) расходы на использование компьютеров для подготовки эталонной выборки и отбора *первичных единиц выборки*;

d) расходы на персонал, занятый отбором *первичных единиц выборки* (если это не делается с помощью компьютера);

e) расходы по работам на местах по составлению списков в единицах выборки предпоследнего этапа, включая расходы на персонал и составление таких материалов, как досье по кластерам;

f) расходы на персонал по сбору данных из выборки домохозяйств в рамках кластеров выборки.

32. В дополнение к расходам обследования на формирование выборки, необходимо вести отчетность по расходам на проведение самого обследования. Такие расходы могут включать: заработную плату регистраторов и контролеров; суточные и командировочные расходы в ходе работ на местах для штатных сотрудников организации, проводящей обследование; транспортные расходы; расходы на топливо; канцелярские принадлежности; расходы на подготовку персонала; расходы на услуги связи и обработку данных.

5.8. Оценка: ограничения данных обследования

33. Значительная часть рассмотренной выше документации по ведению надлежащей отчетности в дополнение к тому, что это важный фактор для обработки результатов обследования, применима также для оценки различных аспектов плана выборки и проведения обследования. Информация по величине долей, ответивших на вопросы, например, помогает оценить, насколько серьезной является погрешность показателей в результате неполучения ответов. Информация по расходам на формирование выборки может использоваться для оценки «экономической» эффективности плана выборки и его полезности для будущих обследований.

34. Как указывалось выше, формальная оценка выборочных обследований охватывает многочисленные аспекты ошибок регистрации данных, которые выходят далеко за пределы настоящего руководства (см. издание Организации Объединенных Наций 1984 года, где приводится исчерпывающий анализ этого вопроса). При этом необходимо отметить, что оценка ошибки регистрации данных должна, например, включать как оперативную деятельность, так и обработку информации. С другой стороны, ошибка выборки поддается оценке, и этот вопрос рассматривается ниже.

35. Несмотря на то что формальные исследования с применением процедур оценки проводятся нечасто, для обследования домашних хозяйств критически важно, тем не менее, чтобы документация обследования содержала информацию об ограничениях данных. Этому вопросу должен быть посвящен краткий раздел в отчетах по существу выводов обследования, зачастую озаглавленный просто «Ограничения данных обследования». В этом разделе читателя следует проинформировать об ошибках обследования, причем применительно как к выборке, так и регистрации данных.

36. В весьма полезной публикации Бюро переписей Соединенных Штатов (1974 год) дается описание того, как представлять информацию об ошибках обследования. В приводимых ниже выдержках из этой публикации (добавление I, стр. I-1) высказывается мнение о том, какого рода информация должна предоставляться пользователям при публикации выводов обследования:

> Приведенная в этом отчете статистика представляет собой оценки, полученные в ходе выборочного обследования. Существуют два возможных типа ошибок в той или иной оценке на основе выборочного обследования — ошибка выборки и ошибка регистрации данных. Ошибки выборки происходят в связи с тем, что измеряется лишь выборка, а не вся совокупность. Ошибки регистрации данных (рассмотренные в главе 8) могут быть отнесены на счет целого ряда источников: неспособность получения информации по всем единицам выборки, трудности по определению понятий, различия в интерпретации вопросов, неспособность или нежелание у части респондентов предоставлять достоверную информацию, ошибки в записи или кодировании полученных данных и прочие ошибки при сборе данных, в полученных ответах, при обработке данных, при оценке величины охвата и процента отсутствующих данных. Ошибки регистрации данных случаются также и в полных переписях.

Точность результата обследования определяется совместным эффектом ошибок выборки и ошибок регистрации данных.

Конкретная выборка, использованная в данном обследовании, — это одна из большого числа всех возможных выборок одинакового размера, которые могли быть отобраны при использовании одного и того же плана выборки. Оценки, полученные из различных выборок, будут отличаться друг от друга. Отклонение выборочной оценки от среднего показателя всех возможных выборок называется ошибкой выборки. Стандартная (среднеквадратическая) ошибка оценки обследования является мерой вариантности оценок, полученных из возможных выборок, а, следовательно, — мерой точности, с которой оценка для конкретной выборки приближается к среднему результату всех возможных выборок. Относительная стандартная ошибка определяется как стандартная ошибка, деленная на измеряемую величину.

Согласно расчетам, приведенным в данном отчете, стандартная ошибка также частично измеряет эффект ошибок регистрации данных, но не измеряет какие-либо систематические погрешности данных. Погрешность — это разница, усредненная по всем возможным выборкам, между оценкой и ожидаемым значением. Очевидно, что правильность результата обследования зависит от ошибок как выборки, так и регистрации данных, измеряемых величиной стандартной ошибки, а также от погрешностей и других видов ошибок регистрации данных, не измеряемых с помощью стандартной ошибки.

37. Вышеизложенное подразумевает, что важным элементом анализа выборки является оценка ошибок выборки, которая должна проводиться в отношении основных оценок обследования. Как обсуждалось выше, одной из отличительных характеристик вероятностной выборки является то, что сама выборка может использоваться для оценки стандартных ошибок. Методы оценки вариантности и стандартной ошибки детально рассматриваются в главе 6. Кроме того, для оценки стандартной ошибки существуют эффективные и надежные программные пакеты, которые следует использовать при малейшей возможности.

38. Как правило, величины стандартных ошибок рассчитываются для ключевых исследуемых характеристик обследования, поскольку непрактично и ненужно делать это для всех позиций обследования. Информация о стандартных ошибках дает пользователям возможность определения надежности оценок обследования и позволяет вывести доверительные интервалы по интересующим их оценкам.

39. Стандартные ошибки могут также использоваться для оценки собственно плана выборки. Особенно полезным статистическим параметром для этой цели является эффект схемы выборки, *deff*, а более точно, *deft*, или квадратный корень из *deff*. Сравнительно просто вычислить величину *deft* для каждого пункта данных, для которого может быть рассчитана величина стандартной ошибки. Для этого нужно всего лишь поделить оценку стандартной ошибки для данного пункта на величину стандартной ошибки для простой случайной выборки такого же размера, а именно: pq/n, где p — расчетная пропорция; $q = 1 - p$, а n — размер выборки. Данный расчет служит для подтверждения или опровержения наличия эффектов схемы, принятых в качестве допущения при составлении плана выборки, поскольку фактические величины *deff* или *deft* будут известны только после завершения обследования, обработки данных и расчета стандартных ошибок.

40. Специалист-статистик по формированию выборки может использовать рассчитанные эффекты схемы, чтобы определить, имеют ли кластеры обоснованный размер для ключевых пунктов данных, и при необходимости — внести соответствующие корректировки. Например, если *deft* значительно превышает ожидаемую величину для ключевых пунктов данных, то план

выборки для будущих обследований может быть составлен таким образом, чтобы использовать более мелкие размеры кластеров.

5.9. Краткие рекомендации

41. В этом разделе в формате контрольного списка приводится резюме основных рекомендаций, содержащихся в данной главе, которые, как и в предыдущих главах, должны восприниматься, скорее, как «эмпирические правила», чем как непреложные рекомендации. Следует:

- документировать связанные с выборкой аспекты обследований двумя путями: надлежащим ведением отчетности и предоставлением технической информации пользователям;
- вести подробную отчетность по процедурам выборки, включая расходы на нее;
- разрабатывать коды для переменных параметров плана выборки, таких как административные районы, *первичные единицы выборки*, кластеры, домохозяйства, лица и т. д.;
- стремиться к стандартизации кодирования переменных параметров выборки, которые единообразны для всех обследований;
- документально фиксировать все отклонения или отступления от первоначального плана выборки, которые случаются при проведении обследования;
- рассчитывать и документально фиксировать уровни вероятностей отбора на каждом этапе формирования выборки;
- в частности, фиксировать информацию по подвыборкам, которые осуществляются в ходе работ на местах;
- документально фиксировать информацию о числе и категориях случаев неполучения ответов;
- документально фиксировать веса схемы, поправки на неполучение ответов и поправки на постстратификацию;
- вести подробную отчетность о расходах по каждой операции в плане выборки и при проведении обследования;
- применительно к эталонной выборке вести отчетность о расходах, как по ее составлению (первоначальному), так и обновлению;
- готовить технические отчеты для пользователей по методологии формирования выборки и проведения обследования;
- готовить краткий отчет по ограничениям данных для всех независимых публикаций по результатам обследования;
- готовить более подробный технический отчет по всем аспектам методологии формирования выборки;
- делать расчет ошибок выборки для основных переменных параметров и включать эти данные в технические отчеты;
- делать расчет эффектов схемы (*deff* или *deft*) для основных переменных параметров;
- назначать отдельного сотрудника, ответственного за подготовку документации.

Справочная литература и дополнительные источники информации

Casley, D. J., and D. A. Lury, (1981). *Data Collection in Developing Countries*. Oxford, United Kingdom: Clarendon Press.

International Statistical Institute (1975). *Manual on Sample Design*. World Fertility Survey Basic Documentation, Voorburg, Netherlands.

League of Arab States (1990). *Sampling Manual, Arab Maternal and Child Health Survey*, Basic Documentation 5. Cairo: Pan Arab Project for Child Development (PAPCHILD).

Macro International, Inc. (1996), *Sampling Manual*. DHS-III Basic Documentation No. 6. Calverton, Maryland.

United Nations (1964). Recommendations for the Preparation of Sample Survey Reports (Provisional Issue). Statistical Papers, Series C, No. 1, Rev.2.

Организация Объединенных Наций (1984), *Руководство по обследованию домашних хозяйств* (Переработанное издание). Методологические исследования, № 31. В продаже под № R.83. XVII.13.

United States Bureau of the Census (1974). Standards for Discussion and Presentation of Errors in Data. Technical Paper 32. Washington, D.C.: United States Bureau of the Census.

_____(1978). *Current Population Survey Design and Methodology*. Technical Paper 40. Washington, D.C.: Bureau of the Census.

World Bank (1999). *Core Welfare Indicators Questionnaire (CWIQ) Handbook*. Washington, D.C.: World Bank.

Глава 6

Построение и использование весов выборки

6.1. Введение

1. В данной главе обсуждаются различные этапы разработки весов выборки и их использования при расчетах оценок исследуемых показателей на основе данных обследования домашних хозяйств. В частности, рассматривается коррекция весов выборки в целях компенсации различных дефектов в сформированной выборке. Обсуждение ограничивается описательными оценками, которые широко представлены в большинстве отчетов обследований. Важнейшие из приведенных здесь тезисов иллюстрируются реальными примерами из текущих обследований, проводимых в развивающихся странах, или примерами, в которых воспроизводятся реальные ситуации, возникающие в ходе обследований.

6.2. Необходимость в весах выборки

2. Обследования домашних хозяйств, как правило, базируются на комплексных планах выборки в основном для сдерживания расходов. Получающиеся в результате выборки чаще всего имеют те или иные дефекты, которые могут привести к возникновению погрешностей или других расхождений между выборкой и эталонной совокупностью. К таким дефектам относятся отбор единиц с неравными вероятностями, неполный охват населения и неполучение ответов. Веса выборки необходимы для исправления этих дефектов и получения, таким образом, надлежащих оценок интересующих показателей. Кратко цели взвешивания таковы:

a) компенсация неравных вероятностей отбора;

b) компенсация на случай неполучения ответов (от единицы);

c) коррекция распределения взвешенной выборки для ключевых искомых переменных показателей (например, возраст, расовая принадлежность и пол) для ее соответствия распределению для известного контингента населения.

3. Процедуры, используемые для любого из этих сценариев, подробно рассматриваются в последующих разделах. После компенсации дефектов выборки веса могут использоваться в оценке представляющих интерес характеристик совокупности, а также в оценке ошибок выборки для получаемых оценок обследования.

4. Когда веса не используются для компенсации дифференцированных долей отбора в пределах страт (если это предусмотрено планом выборки), а также указанных выше дефектов выборки, получающиеся в результате оценки параметров совокупности, как правило, имеют искажения. См. в разделах 6.3, 6.4 и 6.5 примеры процедур взвешивания, применяемых для каждого из сценариев, включая сравнение в каждом случае взвешенных и невзвешенных оценок.

6.2.1. Обзор

5. Раздел 6.3 посвящен разработке весов выборки в контексте многоэтапного плана выборки, включая коррекцию весов выборки для учета дублирования единиц выборки, а также учета тех единиц, для которых на момент формирования выборки неизвестно соответствие критериям обследования. В разделе 6.4 обсуждаются вопросы взвешивания неравных вероятностей отбора; в нем дается несколько численных примеров, включая рассмотрение конкретного случая разработки весов для национального обследования домашних хозяйств; в заключительной части рассматриваются самовзвешенные выборки. Проблемы неполучения ответов и неполного охвата в обследованиях домашних хозяйств рассматриваются в разделах 6.5 и 6.6, соответственно. Обсуждаются источники и последствия ситуации с неполучением ответов и неполным охватом, а также представлены методы компенсации для случаев неполучения ответов и неполного охвата, включая численные примеры, иллюстрирующие коррекцию весов выборки применительно к фактам неполучения ответов и неполного охвата. Раздел 6.7 посвящен проблеме увеличения вариантности оценок обследования в результате использования весов выборки при анализе данных обследования домашних хозяйств. Представлен также численный пример для иллюстрации расчета касательно увеличения вариантности по причине взвешивания. В разделе 6.8 обсуждается проблема выравнивания весов и представлен пример процедуры выравнивания, в котором выравниваемые веса масштабируются таким образом, чтобы они в итоге давали сумму первоначальных весов. Некоторые заключительные соображения изложены в разделе 6.9.

6.3. Разработка весов выборки

6. Разработку весов выборки можно начинать после определения вероятностей отбора единиц выборки. Вероятность отбора единицы выборки зависит от плана выборки, используемого для отбора данной единицы. В главе 3 дано подробное описание наиболее широко используемых планов выборки и соответствующих этим планам вероятностей отбора. При этом неизменно делается допущение о том, что вероятности отбора имеют определенную величину.

7. Разработка весов выборки иногда рассматривается как первый шаг в анализе данных обследования. Процедура обычно начинается с построения *базисного веса* или *веса схемы* для каждой единицы выборки, с тем чтобы отразить их неравные вероятности отбора. Базисный вес единицы выборки является обратной величиной вероятности ее отбора для включения в выборку. В математическом выражении, если единица включается в выборку с вероятностью p_i, тогда ее базисный вес, обозначаемый как w_i, рассчитывается по формуле:

$$w_i = 1/p_i . \qquad (6.1)$$

8. Например, единица выборки, отобранная с вероятностью 1/50, представляет 50 единиц совокупности, из которых была сформирована выборка. Следовательно, веса выборки выступают в качестве повышающих коэффициентов, предназначенных для обозначения числа единиц в совокупности обследования, представленных единицей выборки, которой придается данный вес. Сумма весов выборки дает несмещенную оценку общего числа единиц в обследуемой совокупности.

9. Применительно к многоэтапным планам базисные веса должны отражать вероятности отбора на каждом этапе. Например, в случае двухэтапной выборки, в которой *i*-ая ПЕВ отбирается с вероятностью p_i на первом этапе выборки, а *j*-ое домохозяйство отбирается в рамках вошедшей в выборку ПЕВ с вероятностью $p_{j(i)}$ на втором этапе выборки, общая вероятность отбора (p_{ij}) каждого домохозяйства выборки является произведением этих двух вероятностей или

$$p_{ij} = p_i \times p_{j(i)}, \tag{6.2}$$

при этом общий базисный вес домохозяйства является, как указывалось ранее, обратным числом от ее общей вероятности отбора. Соответственно, если базисный вес j-го домохозяйства равен $w_{ij,b}$, вес, обусловленный компенсацией на случай неполучения ответов, составляет $w_{ij,nr,}$ а вес, обусловленный компенсацией на случай неполного охвата, равен $w_{ij,nc}$, тогда общий вес данного домохозяйства определяется по формуле:

$$w_{ij} = w_{ij,b} \times w_{ij,nr} \times w_{ij,nc} . \tag{6.3}$$

6.3.1. Коррекция весов выборки при неизвестной степени соответствия критерию отбора

10. Во время сбора данных в ходе обследований домашних хозяйств иногда бывают случаи, когда под вопрос ставится соответствие того или иного домохозяйства критериям обследования. Например, регистратор может никого не обнаружить в жилой единице на момент сбора данных или при повторных посещениях. В этом случае неизвестно, заселена данная жилая единица или нет. Если она заселена, тогда ее надлежит классифицировать как не ответившую жилую единицу (по категории «отсутствие жильцов дома»). В противном случае, такая единица выходит за рамки обследования, и поэтому ее нельзя засчитывать в качестве единицы выборки. Иногда регистраторы исходят из того, что если никого не обнаружено в жилой единице и в ходе повторных посещений, то данная жилая единица не заселена и, следовательно, не может включаться в выборку. Как правило, такое предположение является неверным, и это часто ведет к ошибочно завышенным долям получения ответов.

11. Когда соответствие каких-либо отобранных жилых единиц выборки критериям обследования неизвестно, их веса необходимо соответствующим образом скорректировать, чтобы учесть этот факт. Идея состоит в том, чтобы сделать определенные допущения, которые позволят оценить долю жилых единиц с неизвестными показателями соответствия критериям, как фактически соответствующие этим критериям. Наиболее простым подходом является взять известное соотношение жилых единиц выборки, соответствующих и несоответствующих критериям, и применить это соотношение к единицам, о которых неизвестно их соответствие критериям. Предположим, например, что выборка из 300 жилых единиц имеет распределение категорий ответов, приведенное в таблице 6.1.

Таблица 6.1.
Категории ответов в обследовании

Категория ответа	Число жилых единиц
Завершенные опросы	215
Число лиц, удовлетворяющих критериям отбора, которые не ответили на вопрос	25
Число лиц, не удовлетворяющих критериям отбора	10
Соответствие критериям отбора неизвестно	50

12. Следует отметить, что пропорция жилых единиц с известной степенью соответствия критериям, которые фактически им удовлетворяют, составляет (215+25)/(215+25+10) = 0,96. Следовательно, мы можем допустить, что такая же доля (0,96) жилых единиц, с неизвестной степенью соответствия критериям, может считаться как удовлетворяющая этим критериям. Иными словами, 96 процентов из 50 жилых единиц с неизвестной степенью соответствия критериям (или 48 жилых единиц), фактически им удовлетворяют. В таком случае мы можем скорректировать веса жилых единиц, соответствующих критериям отбора (по категориям «завершенные опросы» и «лица, соответствующие критериям отбора, не ответившие на вопрос»), используя следующий поправочный коэффициент:

$$F_{ue} = \frac{\sum_{c} w_{ij,b} + \sum_{nr} w_{ij,b} + \varepsilon \times \sum_{ue} w_{ij,b}}{\sum_{c} w_{ij,b} + \sum_{nr} w_{ij,b}}, \qquad (6.4)$$

где ε обозначает долю жилых единиц с неизвестной степенью соответствия, оцениваемых как удовлетворяющие критериям (в данном примере $\varepsilon = 0{,}96$). Слагаемые компоненты над символами *c*, *nr* и *ue* в приведенной выше формуле обозначают, соответственно, сумму базисных весов жилых единиц с проведенными опросами: единиц, соответствующих критериям, но не ответивших, и единиц с неизвестной степенью соответствия критериям. Затем получаем скорректированные базисные веса жилых единиц с проведенными опросами и единиц, соответствующих критериям, но не ответивших, путем умножения их первоначальных базисных весов $w_{ij,b}$ на коэффициент F_{ue}.

6.3.2. Коррекция весов выборки на случаи дублирования в выборках

13. Если заранее известно, что некоторые единицы дважды продублированы в инструментарии, тогда повышенная вероятность отбора таких единиц может быть компенсирована за счет придания им весовых коэффициентов, которые являются обратными величинами от числа продублированных записей в списках инструментария, если такие единицы попадают в выборку. Однако достаточно часто дублирование записей обнаруживается только после завершения формирования выборки, и вероятности отбора таких вошедших в выборку единиц должно быть скорректировано для учета этого дублирования. Такая коррекция проводится следующим образом. Предположим, что вошедшая в выборку *i*-ая единица имеет вероятность отбора, обозначенную p_{i1}, и предположим, что существует *k*-1 дополнительных записей в инструментарии выборки, которые выявлены как дублирующие данную единицу выборки, при этом каждый из таких дубликатов имеет вероятности выборки, обозначаемые $p_{i2}, \ldots, p_{ik}$. Тогда скорректированная вероятность отбора данной единицы выборки определяется по формуле:

$$p_i = 1 - (1 - p_{i1})(1 - p_{i2}) \ \ldots \ (1 - p_{ik}). \qquad (6.5)$$

Затем данная единица выборки соответствующим образом взвешивается, а именно с использованием коэффициента $1/p_i$.

14. Процедуры построения весов выборки по приведенным выше сценариям проиллюстрированы далее на конкретных примерах.

6.4. Взвешивание для случая неравных вероятностей отбора

15. Для простоты изложения рассмотрим двухэтапный план выборки, в котором в качестве ПЕВ выступают счетные участки переписи, а в качестве единиц второго этапа — домохозяйства. Предположим, что равновероятностная выборка из n ПЕВ отобрана из общего числа N на первом этапе выборки, а затем m домохозяйств отобраны из каждой вошедшей в выборку ПЕВ. Очевидно, что вероятность отбора домохозяйств будет зависеть от общего числа домохозяйств в той ПЕВ, в которой они находятся. Пусть M_i обозначает число домохозяйств в ПЕВ$_i$. Тогда вероятность отбора ПЕВ составит n/N, а условная вероятность отбора домохозяйства в i-ой вошедшей в выборку ПЕВ, составит m/M_i. Таким образом, общая вероятность отбора домохозяйства определяется уравнением:

$$p_{ij} = p_i \times p_{j(i)} = \frac{n}{N} \times \frac{m}{M_i} = \frac{nm}{N} \times \frac{1}{M_i} . \qquad (6.6)$$

Кроме того, вес включенного в выборку домохозяйства по данному плану определяется как:

$$w_i = \frac{1}{p_{ij}} = \frac{N}{nm} \times M_i . \qquad (6.7)$$

Пример 1

Равновероятностная выборка из 5 домохозяйств отобрана из 250 домохозяйств. Один взрослый человек отбирается в случайном порядке из каждого вошедшего в выборку домохозяйства. Записывается месячный доход (y_{ij}) и уровень образования ($z_{ij} = 1$, если среднее и высшее образование; 0 — в остальных случаях) j-го включенного в выборку взрослого человека в i-ом домохозяйстве. Пусть M_i обозначает число взрослых членов i-го домохозяйства. Тогда общая вероятность отбора того или иного взрослого человека в выборку определяется как

$$p_{ij} = p_i \times p_{j(i)} = \frac{5}{250} \times \frac{1}{M_i} = \frac{1}{50} \times \frac{1}{M_i} .$$

Следовательно, весовое значение включенного в выборку взрослого человека определяется как:

$$w_i = \frac{1}{p_{ij}} = 50 \times M_i .$$

16. Теперь мы проиллюстрируем расчет базисных оценок согласно указанному выше плану выборки. Предположим, что данные, полученные от одного включенного в выборку взрослого человека для каждого домохозяйства, которое вошло в сформированную на первом этапе выборку из пяти домохозяйств, представлены в таблице 6.2, ниже. Отметим, что данные о числе взрослых лиц в каждом домохозяйстве и соответствующем общем весовом значении включенного в выборку из каждого домохозяйства человека приводятся во втором и третьем столбцах, соответственно.

Таблица 6.2
Весовые значения при неравновероятностном отборе

Домохозяйство выборки	M_i	w_i	y_{ij}	z_{ij}	$w_i y_{ij}$	$w_i z_{ij}$	$w_i z_{ij} y_{ij}$
1	3	150	70	1	10 500	150	10 500
2	1	50	30	0	1 500	0	0
3	3	150	90	1	13 500	150	13 500
4	5	250	50	1	12 500	250	12 500
5	4	200	60	0	12 000	0	0
Всего	**16**	**800**	**300**	**3**	**50 000**	**550**	**36 500**

17. Оценки различных характеристик могут тогда быть получены из таблицы 6.2 следующим образом:

Расчетная величина среднемесячного дохода определяется как:

$$\bar{y}_w = \frac{\sum w_i y_{ij}}{\sum w_i} = \frac{50\,000}{800} = 62{,}5;$$

Если не используются веса, эта величина составит 60 (или 300/5).

Оценка доли лиц со средним или высшим образованием составляет:

$$\bar{y}_w = \frac{\sum w_i z_{ij}}{\sum w_i} = \frac{550}{800} = 0{,}6875 \text{ или } 68{,}75\% \text{ ;}$$

Если не используются веса, эта оценка составит 3/5 или 0,60, или 60 процентов.

Оценка общего числа взрослых лиц со средним или высшим образованием составит:

$$\hat{t} = \sum w_i z_{ij} = 550.$$

Расчетная величина среднемесячного дохода взрослых лиц со средним или высшим образованием составляет:

$$\bar{y}_w = \frac{\sum w_i z_{ij} y_{ij}}{\sum w_i z_{ij}} = \frac{36\,500}{550} = 66{,}36.$$

18. Иногда веса выборки «нормируются», иными словами, веса умножаются на величину отношения

$$\frac{\text{число респондентов}}{\text{сумма весов всех респондентов}} \cdot \qquad (6.8)$$

19. Следовательно, суммой нормированных весов является реализованный размер выборки для анализа (число респондентов). Отметим, что нормированные веса нельзя использовать для оценки суммарных величин, таких как общее число взрослых лиц со средним или высшим образова-

нием. В этом случае необходимо взвешивать включенные в выборку единицы путем умножения на обратную величину вероятности их отбора, иными словами, необходимо использовать обычные веса выборки. Однако для оценки средних значений и долей необходимо, чтобы веса были лишь пропорциональными обратным величинам вероятностей отбора. Иными словами, неважно, используются ли обычные веса или нормированные веса (пропорциональные обычным весам) для получения оценок средних значений параметров населения, таких как средняя численность или доля женщин детородного возраста, имеющих доступ к первичной медицинской помощи. Оба типа весов дадут один и тот же результат.

20. Например, в предыдущем примере веса w_i пропорциональны M_i (w_i=50× M_i). Если мы используем M_i в качестве весов, тогда оценка доли взрослых лиц со средним или высшим образованием составит:

$$\hat{p} = \frac{\sum M_i z_{ij}}{\sum M_i} \frac{3\times1+1\times0+3\times1+5\times1+4\times0}{3+1+3+5+4} = \frac{11}{16} = 0{,}6875$$

или 68,75 процента, т. е. то же самое, что и ранее. Однако для оценки общего числа взрослых лиц со средним или высшим образованием необходимо использовать обычные веса выборки (w_i = 50× M_i), с тем чтобы получить правильный результат, а именно:

$$\hat{t}_s = \sum(50\times M_i)z_{ij} = 50\sum M_i z_{ij} = 50\times11 = 550.$$

Двухэтапная выборка домохозяйств формируется в сельских районах страны. На первом этапе в выборку включаются 50 деревень с вероятностью, пропорциональной числу домохозяйств во время последней переписи. Общее число домохозяйств в сельских районах во время последней переписи составляло 300 000. После первого этапа выборки следует операция по составлению списков, жилых единиц в каждой из вошедших в выборку деревень. Иногда обнаруживалось, что в одной жилой единице проживало несколько домохозяйств.

21. А теперь рассмотрим различные варианты планов подвыборки (для отбора домохозяйств из вошедших в выборку жилых единиц) и выведем уравнение отбора для общей вероятности отбора домохозяйства для включения в выборку. Пусть D_i обозначает число жилых единиц в деревне i и пусть H_{ij} обозначает число домохозяйств в жилой единице j деревни i. Общее число домохозяйств в деревне обозначается как H_i, и в этом случае определяется уравнением:

$H_i = \sum_j H_{ij}$. Отметим, что $\sum_i H_i = \sum_i \sum_j H_{ij} = 300\,000$.

Рассчитанные таким способом вероятности отбора базируются на формулах, приведенных в главе 3.

6.4.2.1. Вариант 1 плана выборки

22. Из списка по каждой включенной в выборку деревне отбираются пятнадцать жилых единиц путем простой случайной выборки без замещения (ПСВБЗ). Все домохозяйства в отобранных жилых единицах включаются в выборку; следовательно, имеются только два этапа формирования выборки — отбор деревень и отбор жилых единиц. По данному плану уравнение отбора для полной вероятности отбора домохозяйства с целью его включения в выборку выражается следующим образом:

p_{ij} = *pr (выбрана деревня i)*× *pr (выбрана жилая единица j при условии отбора i)*.

Тогда:

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{15}{D_i} = \frac{750}{\sum_i M_i} \times \frac{H_i}{D_i},$$

при этом базисный вес определяется уравнением:

$$w_{ij} = \frac{\sum_i M_i}{750} \times \frac{D_i}{H_i}.$$

23. Отметим, что полная вероятность отбора может варьироваться от деревни к деревне в зависимости от соотношения числа домохозяйств к числу жилых единиц, H_i/D_i. Таким образом, мы приходим к выводу, что данный план не является самовзвешенным (для более подробного рассмотрения самовзвешенных планов выборки см. раздел 6.4.2, ниже). Он был бы самовзвешенным, если бы каждая жилая единица содержала только по одному домохозяйству, т. е. если бы соотношение H_i/D_i было одинаковым для всех вошедших в выборку деревень.

6.4.2.2. Вариант 2 плана выборки

24. Производится систематическая выборка жилых единиц в рамках вошедших в выборку деревень, при этом доля выборки в каждой деревне обратно пропорциональна числу домохозяйств, входивших в нее во время последней переписи. Все домохозяйства в отобранных жилых единицах включаются в выборку. Как и в предыдущем случае, существуют только два этапа отбора: отбор деревень и жилых единиц. Условная вероятность отбора жилой единицы во включенной в выборку деревне *i* может быть выражена как k/H_i, где k — это коэффициент пропорциональности. Таким образом, уравнение отбора для полной вероятности отбора домохозяйства для включения в выборку по данному плану выражается следующим образом:

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} = \frac{50 \times k}{\sum_i M_i},$$

при этом базисный вес определяется уравнением

$$w_{ij} = \frac{\sum_i M_i}{50 \times k},$$

что является постоянной величиной. Таким образом, мы приходим к выводу, что 2-й вариант плана является самовзвешенным планом выборки.

6.4.2.3. Вариант 3 плана выборки

25. Производится систематическая выборка жилых единиц в рамках вошедших в выборку деревень, при этом доля выборки в каждой деревне обратно пропорциональна числу домохозяйств, входивших в нее во время последней переписи. Одно домохозяйство отбирается в случайном порядке из каждой отобранной жилой единицы. В этом случае существуют три этапа отбора: деревни, жилые единицы и домашние хозяйства. Таким образом, уравнение отбора для полной вероятности отбора домохозяйства для включения в выборку по данному плану выражается следующим образом:

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} \times \frac{1}{H_{ij}},$$

при этом базисный вес определяется уравнением

$$w_{ij} = \frac{\sum_i M_i}{50} \times \frac{H_{ij}}{k},$$

и данная величина будет меняться для разных жилых единиц в зависимости от числа домохозяйств в той или иной жилой единице. Таким образом, мы приходим к выводу, что 3-й вариант плана не является самовзвешенным.

6.4.1. Исследование конкретных примеров построения весов: Национальное обследование в сфере здравоохранения Вьетнама, 2001 год

26. Теперь мы переходим к иллюстрации построения весов выборки для фактически проведенного во Вьетнаме в 2001 году обследования — Национального обследования в сфере здравоохранения. Это обследование базировалось на стратифицированном трехэтапном плане выборки. Всего было создано 122 страты, ограниченные городскими и сельскими областями обследования в 61 провинции. Затем формирование выборки было проведено независимо в каждой страте. На первом этапе отбирались общины и городские микрорайоны с вероятностью, пропорциональной размеру (исходя из числа домохозяйств во время переписи населения и жилого фонда 1999 года). На втором этапе в каждой включенной в выборку общине или микрорайоне отбирались два счетных участка (СУ) с помощью систематической выборки с долей выборки, обратно пропорциональной числу счетных участков в данной общине или микрорайоне. На третьем и последнем этапе из каждого вошедшего в выборку СУ отбирались 15 домохозяйств, опять же с помощью систематической выборки.

27. Базисные веса для вошедших в выборку домохозяйств в рамках Национального обследования в сфере здравоохранения могут быть разработаны следующим образом. Допустим, что H_i и E_i обозначают, соответственно, число домохозяйств и число СУ (во время переписи 1999 года) в общине i, и пусть H_{ij} обозначает число домохозяйств в счетном участке j общины i. Тогда полная вероятность отбора домохозяйства k в счетном участке j общины i определяется уравнением

$$p_{ijk} = n_c \times \frac{H_i}{\sum_i H_i} \times \frac{2}{E_i} \times \frac{15}{H_{ij}},$$

где n_c — это число общин, отобранных в данной страте, а

$\sum_i H_i$ — это общее число домохозяйств в данной страте.

Вес выборки домохозяйства (w_{ijk}) является обратной величиной вероятности отбора, а именно:

$$w_{ijk} = \frac{E_i \times H_{ij} \times \sum_i M_i}{30 \times n_c \times H_i}. \qquad (6.9)$$

6.4.2. Самовзвешенные выборки

28. Когда веса всех включенных в выборку единиц равны, такая выборка называется *самовзвешенной*. Даже несмотря на то что единицы более высокого этапа часто отбираются с различными показателями вероятности для повышения эффективности выборки, такие переменные вероятности могут уравновешиваться вероятностями отбора на последующих этапах. 2-й вариант плана выборки в приведенном выше примере 2 иллюстрирует именно такую ситуацию.

29. Однако на практике выборки обследований домашних хозяйств очень редко бывают самовзвешенными на национальном уровне по целому ряду причин. Во-первых, единицы выборки зачастую отбираются с неравной вероятностью в соответствии с планом. Действительно, даже несмотря на то, что ПЕВ нередко отбираются с вероятностью, пропорциональной размеру, а домохозяйства в рамках ПЕВ отбираются по соответствующим показателям, чтобы обеспечить самовзвешенный план выборки, все это может быть сведено на нет за счет отбора для опроса одного лица в каждом включенном в выборку домохозяйстве. Во-вторых, сформированная выборка часто имеет погрешности, включая неполучение ответов (раздел 6.5) и неполный охват (раздел 6.6). В-третьих, необходимость получения точных оценок по областям обследования и особым подгруппам населения зачастую требует осуществления дополнительной выборки по таким областям, с тем чтобы получить достаточно большие размеры выборки для удовлетворения заранее оговоренных требований точности. В-четвертых, когда план выборки предусматривает составление обновленных списков домохозяйств в кластерах выборки (первичных единицах выборки или единицах выборки второго этапа) и требует отбора в каждом кластере заранее определенного фиксированного числа домохозяйств, фактическая вероятность отбора домохозяйства будет в некоторой степени отличаться от своей предусмотренной планом вероятности, которая базировалась на данных инструментария, а не на текущем подсчете домохозяйств; вследствие этого неравные вероятности отбора возникнут даже в том случае, если планировалось достижение самовзвешенного плана выборки.

30. Несмотря на указанные сдерживающие факторы, получение самовзвешенных выборок должно быть целью любой работы по составлению плана выборки в силу преимуществ, которые предлагают такие выборки в отношении как осуществления плана выборки, так и анализа данных, полученных на основе этого плана. Используя самовзвешенные выборки, оценки обследования можно получать из невзвешенных данных, а затем, при необходимости, результаты могут увеличиваться на постоянный коэффициент для получения надлежащих оценок параметров населения. Более того, проще делать анализ, основанный на самовзвешенных выборках, и его результаты могут быть легче поняты и восприняты людьми без знания статистики и широкой общественностью.

6.5. Корректировка весов выборки в случаях неполучения ответов

31. В ходе обследований редко удается получить всю необходимую информацию от всех единиц выборки. Например, некоторые домохозяйства могут вообще не предоставить данных, в то время как другие домохозяйства могут обеспечить лишь частичные данные, т. е. данные по некоторым, но не по всем вопросам обследования. Первый тип неполучения ответов называется *неполучением ответов от единицы* или *полным отсутствием ответов*, тогда как последний называется *неполучением ответов на отдельные вопросы*. При наличии каких-либо систематических расхождений между респондентами и нереспондентами любые первичные оценки, основанные исключительно на ответах респондентов, будут искаженными. Ключевым аспектом надлежащей практики проведения обследований, который подчеркивается по всему тексту настоящего

руководства, является то, насколько важно сохранять долю неполучения ответов в ходе обследования на возможно низком уровне. Это необходимо для снижения возможности внесения тех или иных погрешностей в оценки обследования за счет невключения в него (или включения непропорционально малой доли) определенной части населения. Например, жители городов, имеющие сравнительно высокий доход, с меньшей вероятностью примут участие в многоцелевом обследовании, включающем блок вопросов о доходах. Неполучение ответов от крупного сегмента этой части населения может повлиять на национальные оценки среднего дохода домохозяйств, уровня образования, грамотности и т. д.

6.5.1. Снижение уровней погрешности в связи с неполучением ответов при обследованиях домашних хозяйств

32. Размер погрешностей, связанных с неполучением ответов, например для выборочного среднего значения, определяется двумя факторами:

- доля населения, не представившего ответы на вопросы;
- величина разности между средним значением исследуемой характеристики в группах, ответивших и не ответивших на вопросы обследования.

33. Вследствие этого снижение уровня погрешностей, связанных с неполучением ответов, требует либо малой доли неполучения ответов, либо незначительных различий между домохозяйствами/лицами, ответившими и не ответившими на вопросы обследования. С помощью надлежащего ведения отчетности по каждой включенной в выборку единице существует возможность, базируясь напрямую на данных обследования, определить долю неполучения ответов для всей выборки в целом, а также для исследуемых подобластей. Более того, можно провести специальные, тщательно спланированные исследования для оценки различий между респондентами и нереспондентами (Groves and Couper, 1998).

34. В групповых обследованиях (в которых данные собираются из одной и той же группы включенных в выборку единиц регулярно в динамике по времени) разработчик обследования имеет доступ к большему объему данных для изучения и внесения поправок для учета влияния погрешностей в связи с возможными случаями неполучения ответов, чем в одноразовых или перекрестных обследованиях. В этом случае неполучение ответов может возникать в связи с потерей тех или иных единиц в ходе обследования или отказом участвовать в последующих раундах обследования из-за усталости респондентов или иными факторами и т. д. Данные, собранные в ходе предыдущих раундов групповых обследований, могут впоследствии использоваться для получения большей информации о различиях между респондентами и нереспондентами и могут служить в качестве основы для методов внесения поправок, которые описываются ниже. Более детальная информация о различных методологиях, которые используются для компенсации в случаях неполучения ответов при исследовании результатов обследований, представлена в работах Брика и Калтона (1996 год) и Лепковски (2003 год) и упоминается в справочной литературе.

6.5.2. Компенсация в случаях неполучения ответа

35. Целый ряд методов можно использовать для увеличения доли ответивших на вопросы и, следовательно, снижения уровня погрешностей, связанных с неполучением ответов в обследованиях домашних хозяйств. Один из методов — это убеждение лиц, отказавшихся отвечать на вопросы, с помощью «повторных посещений», в ходе которых регистраторы делают не одну, а несколько попыток завершить опрос включенного в выборку домохозяйства. Более высокие

показатели получения ответов могут быть также достигнуты с помощью улучшения подготовки регистраторов. Тем не менее, независимо от того, сколько усилий направлено на повышение доли домохозяйств, ответивших на вопросы, неполучение ответов будет всегда оставаться неизбежным признаком любого обследования домохозяйств. Вследствие этого разработчики обследований часто вносят поправки для компенсации в случаях неполучения ответов. Существуют два следующих основных подхода к корректировкам на случай неполучения ответов от единицы:

a) корректировка размера выборки путем формирования более крупного, чем это необходимо, первоначального размера выборки для компенсации ожидаемых случаев неполучения ответов;

b) корректировка весов выборки для учета случаев неполучения ответов.

36. Рекомендуется всегда решать проблему неполучения ответов от единицы выборки в обследованиях домашних хозяйств с помощью корректировки весов выборки для учета отказавшихся от ответов домохозяйств. В разделе 6.5.3 дается краткий обзор шагов по осуществлению корректировки весов выборки в случаях неполучения ответов, и приводится математический пример.

37. Существует несколько проблем, связанных с заменой, которая эквивалентна вменению полного файла данных по не ответившей единице выборки (Kalton, 1983). Во-первых, это увеличивает вероятность отбора возможных субститутов, поскольку не вошедшие в выборку домохозяйства, близкие к не ответившим, но вошедшим в выборку домохозяйствам, имеют более высокую вероятность отбора, чем близкие к ответившим и вошедшим в выборку домохозяйствам. Во-вторых, усилия по замене не ответивших домохозяйств отнимают много времени, подвержены ошибкам и искажениям и трудно поддаются проверке и контролю. Например, замена может производиться, используя «удобное» домохозяйство, а не то домохозяйство, которое конкретно предназначено, для того чтобы служить субститутом или заменой не ответившего домохозяйства, что, таким образом, служит еще одним источником погрешности. В силу всех этих проблем замена не должна производиться для компенсации в случаях неполучения ответов в обследованиях домашних хозяйств, если не существует надлежащего обоснования для ее использования в конкретных целях.

38. Для случаев неполучения ответов на отдельные вопросы стандартным методом компенсации является *вменение значений*, которое не рассматривается в данном руководстве.

6.5.3. Корректировка весов выборки в случаях неполучения ответов

39. Процедура корректировки весов выборки часто используется в крупномасштабных обследованиях домашних хозяйств для компенсации в случаях неполучения ответов. По сути, эта корректировка переносит базисные веса всех удовлетворяющих критериям отбора и включенных в выборку, но не ответивших единиц выборки на ответившие единицы и сводится к следующим шагам:

- *Шаг 1.* Применение первоначальных весов схемы (к неравным вероятностям отбора и, если уместно, к другим корректировкам, рассмотренным в предшествующих разделах).
- *Шаг 2.* Разделение выборки на подгруппы и расчет взвешенных долей получения ответов для каждой подгруппы.
- *Шаг 3.* Использование обратных величин долей ответивших на вопросы в данной подгруппе для корректировок в случаях неполучения ответов.
- *Шаг 4.* Расчет скорректированного веса в случае неполучения ответов для *i*-ой единицы выборки с помощью следующего уравнения:

$$w_i = w_{1i} \times w_{2i}, \tag{6.10}$$

где w_{1i} — это первоначальный вес, а w_{2i} — скорректированный вес в случае неполучения ответов. Отметим, что взвешенная доля не ответивших на вопросы может определяться как отношение взвешенного числа проведенных опросов для удовлетворяющих критериям отбора и включенных в выборку случаям к взвешенному числу удовлетворяющих критериям отбора и включенных в выборку случаев.

Пример

Стратифицированная многоэтапная выборка из 1 000 домохозяйств отбирается по двум регионам страны (северному и южному). Домохозяйства на севере отбираются с долей выборки 1/100, а домохозяйства на юге — с долей 1/200. Доля получения ответов в городских районах ниже, чем в сельских. Пусть n_h обозначает число домохозяйств, включенных в выборку в страте h, а r_h обозначает число удовлетворяющих критериям отбора домохозяйств, ответивших на вопросы обследования; допустим также, что t_h обозначает число ответивших домохозяйств, имеющих доступ к первичному медицинскому обслуживанию. Тогда скорректированный вес для не ответивших домохозяйств в страте h определяется как

$$w_h = w_{1h} \times w_{2h}, \tag{6.11}$$

где $w_{2y} = n_h/r_h$. Предположим, что величины показателей на уровне страт приведены в таблице 6.3.

Таблица 6.3
Корректировка при взвешивании в случаях неполучения ответов

Страта	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h r_h$	$w_h t_h$
Север–городские районы	100	80	70	100	1,25	125	10 000	8 750
Север–сельские районы	300	120	100	100	2,50	250	30 000	25 000
Юг–городские районы	200	170	150	200	1,18	236	40 120	35 400
Юг–сельские районы	400	360	180	200	1,11	222	79 920	39 960
Всего	**1 000**	**730**	**500**				**160 040**	**109 110**

Таким образом, расчетная доля домохозяйств, имеющих доступ к первичному медицинскому обслуживанию, составляет:

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109\,110}{160\,040} = 0{,}682 \text{ или } 68{,}2\%,$$

а расчетное число домохозяйств, имеющих доступ к первичному медицинскому обслуживанию, составляет:

$$\hat{t} = \sum w_h t_h = 109\,110 = 68{,}2\% \text{ от } 160\,040.$$

Следует отметить, что невзвешенная расчетная доля домохозяйств, имеющих доступ к первичному медицинскому обслуживанию, основанная лишь на данных респондентов, составляет:

$$\hat{p}_{uw} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0{,}685 \text{ или } 68{,}5\% \ ,$$

а расчетная доля с применением первоначальных весов без корректировки в случаях неполучения ответов составляет:

$$\hat{p}_1 = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83\,000}{126\,000} = 0{,}659 \text{ или } 65{,}9\% \ .$$

40. Этот пример приведен также для того, чтобы проиллюстрировать, каким образом первоначальные веса корректируются для компенсации в случаях неполучения ответов. Результаты показывают значительное расхождение между расчетной долей при использовании только первоначальных весов и расчетной долей с использованием скорректированных весов в случаях неполучения ответов, при этом разница между невзвешенной долей и долей, скорректированной в связи с неполучением ответов, похоже, является незначительной.

41. После корректировки весов в случаях неполучения ответов, при необходимости, могут производиться и дополнительные корректировки весов. В следующем разделе рассматривается корректировка весов в случаях неполного охвата.

6.6. Корректировка весов выборки в случаях неполного охвата

42. Неполный охват обозначает неспособность инструментария выборки охватить всю целевую совокупность, в результате чего у некоторых единиц совокупности отсутствует вероятность отбора для включения в выборку, формируемую для обследования домашних хозяйств. Это лишь один из многих возможных недостатков инструментариев выборки, используемых при формировании выборок для обследований, более подробное обсуждение инструментариев выборки см. в главе 4.

43. Неполный охват является серьезной проблемой при обследованиях домашних хозяйств, особенно тех, которые проводятся в развивающихся странах. О негативном эффекте неполного охвата свидетельствует тот факт, что полученные из выборки оценки численности населения по результатам некоторых обследований в развивающихся странах были значительно меньше оценок населения, полученных из других источников. Вследствие этого методология выявления, оценки и контроля неполного охвата в ходе обследований домашних хозяйств должна быть ключевым направлением работы и подготовки кадров для национальных статистических служб.

44. В данном разделе рассматриваются некоторые причины неполного охвата в обследованиях домашних хозяйств, а также одна из процедур, применяемых для компенсации в случаях неполного охвата, а именно — статистическая корректировка весов методом последующей стратификации.

6.6.1. Причины неполного охвата в обследованиях домашних хозяйств

45. Большинство обследований домашних хозяйств в развивающихся странах базируются на планах многоэтапной стратифицированной территориальной случайной выборки. Единицами первого этапа или первичными единицами выборки обычно являются географические территориальные единицы. На втором этапе составляется список домохозяйств или жилищных единиц, из которого формируется выборка домохозяйств. На последнем этапе составляется список членов или жителей домохозяйств, из которого формируется выборка отдельных лиц. Следовательно, неполный охват может иметь место на любом из трех уровней: уровень ПЕВ, уровень домохозяйств и уровень отдельных лиц.

46. Поскольку ПЕВ, как правило, базируются на счетных участках, определенных и использованных в предыдущей переписи населения и жилого фонда, то ожидается, что они обеспечат полный географический охват исследуемого населения. В связи с этим степень неполного охвата ПЕВ обычно невелика. Для обследований домашних хозяйств в развивающихся странах неполный охват ПЕВ не является столь же серьезной проблемой, как неполный охват на последующих этапах плана выборки. Тем не менее неполный охват ПЕВ все же имеет место в большинстве обследований. Например, даже если то или иное обследование может быть предназначено для получения оценок по всему населению страны или региона этой страны, определенные ПЕВ могут преднамеренно исключаться на стадии разработки плана выборки, поскольку некоторые регионы страны стали недоступными по причине гражданской войны или волнений, стихийных бедствий или иных факторов. Кроме того, из инструментариев выборки иногда удаляются отдаленные районы с очень небольшим числом домохозяйств или жителей, поскольку их охват связан со слишком высокими расходами. В связи с тем, что они представляют собой лишь небольшую долю населения, их влияние на показатели численности населения весьма незначительно (более подробное обсуждение, включая многочисленные примеры неполного охвата ПЕВ в обследованиях домашних хозяйств, см. в главе 4). В отчете по результатам такого обследования должен быть четко указан факт исключения этих районов. Не должно создаваться впечатление о том, что результаты обследования применимы ко всей стране или региону, когда фактически не охвачена часть населения. Характеристики, касающиеся неполного охвата обследования, должны быть в полной мере указаны в отчете об этом обследовании.

47. Неполный охват становится более серьезной проблемой на уровне домохозяйств. Большинство обследований рассматривают домашнее хозяйство, как группу лиц, которые обычно имеют те или иные родственные связи и которые обычно проживают в жилом помещении или жилищной единице. Необходимо решить некоторые важные проблемы по определению понятий, например, кого надлежит считать постоянным жильцом и что является жилищной единицей. Каким образом следует классифицировать структуры, включающие много жилищных единиц (такие, как многоквартирные дома) и жилищные единицы с несколькими домохозяйствами? Подчас легко выявить жилищные единицы, однако, сложные социальные структуры могут затруднить выявление домохозяйств в рамках какой-либо жилищной единицы. Вследствие этого возникает масса возможностей для неправильной или непоследовательной трактовки таких концепций со стороны разных регистраторов или в различных странах и культурах. В любом случае необходимы строгие рабочие инструкции для правильной ориентировки регистраторов в том плане, кого надлежит считать членом домохозяйства и что считать жилищной единицей.

48. К прочим факторам, способствующим неполному охвату, можно отнести исключение по невнимательности жилищных единиц из списков, подготовленных в ходе работ на местах, или исследуемых подгрупп населения (например, детей младшего возраста или лиц пожилого возраста), пропуски из-за ошибок в измерениях, невключение в список отсутствующих членов домохозяйства и пропуски в связи с неправильным пониманием концепций обследования. Существует также временной аспект этой проблемы, иными словами, жилищные единицы могут оказаться незаселенными или находиться в стадии строительства на момент составления списка, однако стать заселенными на момент сбора данных. Применительно к обследованиям домашних хозяйств в развивающихся странах проблема неполного охвата обостряется за счет того, что проводимые там переписи населения — единственная основа для формирования инструментариев выборки — в большинстве случаев не содержат подробных адресов единиц выборки на уровнях домохозяйств и отдельных лиц. Зачастую используются устаревшие или неточные административные списки домохозяйств, а отдельные люди в рамках домохозяйств преднамеренно или случайно исключа-

ются из списков членов домохозяйств. Более подробно причины неполного охвата рассматриваются в работе Лепковски (2003 год) и приведенных в ней источниках.

6.6.2. Компенсация неполного охвата в обследованиях домашних хозяйств

49. Существует целый ряд подходов к решению проблемы неполного охвата в обследованиях домашних хозяйств (Lepkowski, 2003). К ним относятся:

a) совершенствование соответствующих процедур на местах, например использование нескольких инструментариев и усовершенствование процедур составления списков;

b) компенсация в случаях неполного охвата посредством статистической корректировки весов.

50. Что касается второго подхода, то при наличии надежных контрольных суммарных показателей по всему населению и по оговоренным подгруппам населения можно попытаться скорректировать веса единиц выборки таким образом, чтобы сумма весов совпадала с контрольными суммарными показателями для оговоренных подгрупп населения. Такие подгруппы называются *последующими стратами*, а сама процедура статистической корректировки называется *последующей стратификацией*. Эта процедура компенсирует случаи неполного охвата путем корректировки взвешенного распределения выборки для определенных переменных параметров, с тем чтобы она соответствовала известному распределению совокупности [некоторые примеры из практики методов анализа данных обслуживания с помощью последующей стратификации см. в работе Лехтонена и Пахкинена, (1995 год)]. Ниже дается простой пример по данной проблеме.

Пример

Предположим, что применительно к предыдущему примеру из независимого источника, такого как текущие данные регистра актов гражданского состояния, стало известно, что число домохозяйств составляет 45 025 на севере и 115 800 на юге. Предположим далее, что взвешенные суммарные значения составляют, соответственно, 40 000 и 120 040. Выполняем два следующих шага:

- *Шаг 1.* Рассчитываем коэффициенты последующей стратификации.

  Для северного региона мы имеем: $w_{3h} = \frac{45\,025}{40\,000} = 1{,}126$; и

  для южного региона мы имеем: $w_{3h} = \frac{115\,800}{120\,040} = 0{,}965$.

- *Шаг 2.* Рассчитываем итоговый скорректированный вес: $w_f = w_h \times w_{3h}$.

  Численные результаты приведены в таблице 6.4.

Таблица 6.4
Взвешивание после стратификации для корректировки в случаях неполного охвата

Страта	r_h	t_h	w_h	w_{fh}	$w_{fh}r_h$	$w_{fh}t_h$
Север–городские районы	80	70	125	140,75	11 260	9 852
Север–сельские районы	120	100	250	281,40	33 768	28 140
Юг–городские районы	170	150	236	227,77	38 721	34 166
Юг–сельские районы	360	180	222	214,20	77 112	38 556
Всего	**730**	**500**			**160 861**	**110 714**

Оцениваемая доля домохозяйств, имеющих доступ к первичному медицинскому обслуживанию, составляет:

$$\hat{p}_f = \frac{\sum w_{fh} t_h}{\sum w_{fh}\, r_h} = \frac{110\,714}{160\,861} = 0{,}69 \text{ или } 69\%.$$

51. Отметим, что при использовании весов, скорректированных с помощью последующей стратификации, взвешенная численность выборки по северному и южному регионам составляет, соответственно, 45 024 (11 256+33 768) и 115 821 (38 709+77 112) домохозяйств, причем эти цифры хорошо согласуются с приведенными выше независимыми контрольными итоговыми показателями.

6.7. Увеличение вариантности выборки по причине взвешивания

52. Даже если использование весов при анализе данных обследования имеет тенденцию к уменьшению искажений оценок, этот шаг также может повысить вариантность таких оценок. Чтобы упростить обсуждение данного вопроса, мы рассмотрим стратифицированный одноэтапный план с равновероятностными выборками в рамках страт. В случае, если вариантность страт (иными словами, вариантность единиц в страте) не одинакова в каждой страте, то наличие неравных весов страты по разным стратам (например, весов, обратно пропорциональных вариантности страт) может дать более точные оценки обследования. Тем не менее, если вариантность страт одинакова в каждой страте, тогда наличие неравных весов приведет к более высокой вариантности оценок обследования, чем при наличии равных весов.

53. Эффект от использования весов состоит в повышении вариантности в расчетном среднем значении совокупности на коэффициент

$$L = n \times \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2}\ , \qquad (6.12)$$

где $n = \sum_h n_h$

— это общий реализованный размер выборки, w_h — итоговый вес и n_h — реализованный размер выборки для страты h. Приведенная выше формула может также быть выражена при помощи коэффициента вариантности весов как

$$L = n \times \frac{\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j)\ , \qquad (6.13)$$

где $CV^2(w_j) = \frac{n}{(\sum_j w_j)^2}\left\{\sum_j w_j^2 - \frac{1}{n}(\sum_j w_j)^2\right\} = \frac{\textit{Вариантность весов}}{(\textit{среднее значение весов})^2}$.

Пример

Теперь рассчитаем коэффициент увеличения вариантности, используя данные примера из раздела 6.6.2, с итоговыми весами w_{fh} и реализованными размерами выборки страт r_h (см. таблицу 6.4).

Таблица 6.5
Параметры вариантности страт

Страта	r_h	w_{fh}	$w_{fh}r_h$	$w^2_{fh}r_h$
Север–городские районы	80	140,75	11 260	1 584 845
Север–сельские районы	120	281,40	33 768	9 502 315
Юг–городские районы	170	227,77	38 721	8 819 459
Юг–сельские районы	360	214,20	77 112	16 517 390
Всего	**730**		**160 861**	**36 424 009**

Таким образом, $L = 730 \times \frac{36\ 424\ 009}{(160\ 861)^2} = 1{,}03$.

Иными словами, по причине использования весов имеет место увеличение вариантности в оценках обследования на уровне примерно 3 процентов.

6.8. Выравнивание весов

54. Когда веса рассчитаны и скорректированы для компенсации рассмотренных выше недостатков, рекомендуется изучить распределение скорректированных весов. Чрезмерно большие веса, даже влияющие лишь на малую долю включенных в выборку случаев, могут привести к значительному возрастанию вариантности оценок обследования. Вследствие этого, распространенной практикой является выравнивание предельных значений весов до некоторой максимальной величины, с тем чтобы ограничить ассоциированную вариантность весов (уменьшая, тем самым, вариантность оценок обследования) и в то же время не допустить доминирования небольшого числа включенных в выборку единиц над общей оценкой. Выравнивание весов наиболее часто используется после корректировки весов в случаях неполучения ответов.

55. Хотя процедура выравнивания весов способна уменьшить вариантность оценок, она также вносит погрешности в оценки. В некоторых случаях уменьшение вариантности вследствие выравнивания экстремально больших значений весов может более чем скомпенсировать увеличение образующихся погрешностей, снижая, таким образом, среднеквадратическую ошибку оценок обследования. На практике выравнивание весов должно осуществляться только при обоснованности такого шага, иными словами, когда можно достоверно установить, что погрешности, появляющиеся в результате использования выравненных (в отличие от первоначальных) весов, в меньшей степени воздействуют на общую среднеквадратическую ошибку, чем соответствующее снижение вариантности, достигаемое в результате выравнивания.

56. Для любого стратифицированного плана выборки процесс выравнивания весов должен в идеальном случае осуществляться в рамках каждой страты. Этот процесс начинается с определения верхней границы первоначальных весов, а затем проводится корректировка всего набора весов, с тем чтобы сумма выравненных весов была равна сумме первоначальных весов. Пусть w_{hi} обозначает итоговый вес i-ой единицы в страте h, и допустим также, что w_{hB} обозначает верхнюю

границу весов, определенных для страты *h*. Тогда выравненный вес для *i*-ой единицы, включенной в выборку в страте *h*, может быть рассчитан как:

$$w_{hi(T)} = \begin{cases} w_{hi} & \text{если } w_{hi} < w_{hB} \\ w_{hB} & \text{если } _{hi} \geq w_{hB} \end{cases} . \qquad (6.14)$$

57. Теперь выравненные веса для всей выборки могут быть дополнительно скорректированы, с тем чтобы их сумма точно совпадала с суммой первоначальных весов. Для простоты изложения мы примем постоянные веса в пределах страт и опустим нижний индекс *i* при дальнейшем рассмотрении вопроса. Пусть F_T обозначает отношение суммы первоначальных весов к сумме выравненных весов, иными словами,

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}}, \qquad (6.15)$$

где суммы в данном соотношении взяты по всем стратам и, следовательно, по всем единицам выборки. Если мы определим скорректированный выравненный вес для *h*-ой страты как

$$w^*_{h(T)} = F_{Th} \times w_{(T)}, \qquad (6.16)$$

тогда ясно, что $\sum_h n_h w^*_{h(T)} = \sum_h n_h w_h$, т. е. желаемый результат достигнут.

Приведенный ниже пример призван проиллюстрировать процедуру выравнивания и помочь ее пониманию.

58. Первые два столбца в таблице 6.6, ниже, показывают общее число единиц и, соответственно, итоговый вес в каждой из семи страт. Выбрано значение максимального веса, равное 250, как показано в третьем столбце этой таблицы.

Таблица 6.6
Выравнивание весов

	n_h	w_h	$w_{h(T)}$	$n_h w_h$	$n_h w_{h(T)}$	$n_h w^*_{h(T)}$
	80	140,75	140,75	11 260	11 260	11 823,00
	100	150,25	150,25	15 025	15 025	15 776,25
	125	175,00	175,00	21 875	21 875	22 968,75
	150	200,00	200,00	30 000	30 000	31 500,00
	120	250,00	250,00	30 000	30 000	31 500,00
	120	275,13	250,00	33 015	30 000	31 500,00
	170	285,40	250,00	48 518	42 500	44 625,00
Всего	865			189 693	180 660	189 693,00

Отметим, что в этом случае,

$$F_T = \frac{\sum_i n_h w_{hi}}{\sum_i n_h w_{hi(T)}} = \frac{189\,693}{180\,660} = 1{,}05.$$

Масштаб выравненных весов был изменен таким образом, чтобы они суммировались в первоначальное общее значение $\sum_h n_h w_h = 189\,693$ путем умножения каждого веса на F_T=1,05.

6.9. Заключительные замечания

59. В настоящее время веса выборки стали восприниматься как неотъемлемая часть анализа данных при обследованиях домашних хозяйств как в развивающихся странах, так и в остальных регионах мира. Большинство программ обследований сегодня предполагают использование весов даже в редких ситуациях наличия самовзвешенных выборок (в этом случае веса будут равны единице). В прошлом разработчиками обследований были приложены большие усилия в плане достижения практически неосуществимой цели, а именно формирования самовзвешенной выборки, что делало введение весов ненужным при анализе данных обследования. Общепринятая точка зрения состояла в том, что использование весов сильно усложняло проведение анализов и что для осуществления анализа взвешенных показателей, если и имелись, то весьма ограниченные вычислительные возможности. Тем не менее прогресс компьютерных технологий за последнее десятилетие опровергает этот аргумент. Компьютерное аппаратное и программное обеспечение сегодня доступно с финансовой точки зрения и имеется в наличии во многих развивающихся странах. Кроме того, уже существует множество специализированных программных пакетов, предназначенных конкретно для анализа данных по результатам обследований. Их обзор и сравнение даются в главе 7.

60. Как рассматривалось ранее, использование весов снижает уровень погрешностей, возникающих в связи с несовершенством выборки и обусловленных неполным охватом и неполучением ответов. Неполучение ответов и неполный охват являются различными видами ошибки в связи с невозможностью в рамках соответствующего плана обследования получить информацию от некоторых единиц целевой совокупности. Для обследований домашних хозяйств в развивающихся странах неполный охват является более серьезной проблемой, нежели неполучение ответов. В настоящей главе приведены примеры процедур по разработке и статистической корректировке базисных весов для компенсации некоторых из указанных проблем в ходе обследований домашних хозяйств, а также по применению скорректированных весов для оценки интересующих параметров. Появление быстродействующих компьютеров, а также доступного по цене или бесплатного статистического программного обеспечения должно превратить использование весов в рутинный аспект анализа данных по результатам обследований домашних хозяйств даже в развивающихся странах. Тем не менее, как показано в данной главе, разработка весов выборки усложняет по разным причинам операции в ходе обследования. Например, веса должны рассчитываться для каждого этапа формирования выборки; их необходимо корректировать для компенсации различных недостатков выборки; и, наконец, их необходимо хранить и использовать надлежащим образом в любом последующем анализе. Следовательно, повышенное внимание необходимо уделять разработке операций взвешивания и фактическому расчету весов, которые предполагается использовать в анализе результатов обследования.

Справочная литература и дополнительные источники информации

Brick, J. M., and G. Kalton, (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, vol. 5, pp. 215–238.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons.

Groves, R. M., and M. P. Couper, (1998). *Non-response in Household Interview Surveys.* New York: John Wiley & Sons.

Groves, R.M., and others (2002). *Survey Non-response.* New York: John Wiley & Sons.

D. Kasprzyk, (1986). The treatment of missing survey data. *Survey Methodology*, vol. 12, pp. 1–16.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Ann Arbor, Michigan: Survey Research Center, University of Michigan.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

I. Hess, (1950). On non-coverage of sample dwellings, *Journal of the American Statistical Association*, vol. 53, pp. 509–524.

Lehtonen, R., and E. J. Pahkinen. (1995). *Practical Methods for Design and Analysis of Complex Surveys*, New York: Wiley.

Джеймс Лепковски (2005 год). «Ошибка, связанная с отсутствием наблюдения, при проведении обследований домашних хозяйств в развивающихся странах». См.в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой.* Методологические исследования, № 96. В продаже под № R.05.XVII.6.

Lessler, J. and W. Kalsbeek, (1992). *Nonsampling Error in Surveys*, New York: John Wiley & Sons.

Levy, P. S., and S. Lemeshow. (1999), *Sampling of Populations: Methods and Applications*, 3rd ed. New York: John Wiley & Sons.

Lohr, S. (1999), *Sampling: Design and Analysis,* Duxbury Press, Pacific Grove.

Yansaneh, I. S. (2004), *Overview of Sample Design Issues for Household Surveys in Developing Countries*, in Household Surveys in Developing and Transition Countries: Technical Report, United Nations, New York.

Глава 7

Расчет ошибок выборки для данных по результатам обследований

7.1. Введение

1. В настоящей главе дается краткий обзор разных методов оценки ошибок выборки в данных по результатам обследований домашних хозяйств, полученных с помощью различных планов выборки, начиная от стандартных планов, которые можно найти в любом вводном пособии по теории выборки [например, Кокран (1977 год)], и заканчивая более сложными планами, которые используются в крупномасштабных обследованиях домашних хозяйств. Для стандартных планов выборки формулы представлены наряду с численными примерами для иллюстрации процесса оценивания ошибок выборки, создания доверительных интервалов и расчета эффектов схемы и эффективных размеров выборки. Затем представлены методы процесса оценивания ошибок выборки в более сложных планах. Рассматриваются достоинства и недостатки каждого метода, и приводятся численные примеры для иллюстрации осуществления конкретных процедур. Представленный пример базируется на данных реального обследования и поясняет тот факт, что стандартные статистические программные пакеты дают заниженные значения ошибок выборки в оценках обследования, что ведет к неправильным выводам касательно целевых параметров обследования. Во избежание таких проблем в настоящей главе настоятельно рекомендуется использовать специальные статистические программные пакеты для полного учета сложного характера тех планов выборки, которые широко применяются для обследований домашних хозяйств. Дается сравнительное описание некоторых из таких программных пакетов.

7.1.1. Расчет ошибок выборки для данных по результатам комплексного обследования

2. Аналитические цели хорошо спланированных обследований домашних хозяйств в последнее время расширились далеко за пределы базисных итоговых таблиц или суммарных величин исследуемых параметров. Ныне аналитики заинтересованы также в построении гипотез и их проверке или построении моделей. Например, вместо простой оценки доли населения, живущего в условиях нищеты или имеющего среднее или высшее образование, аналитики сегодня стремятся оценить последствия той или иной политики или изучить, каким образом на ту или иную ключевую переменную величину в ответах, например оценки успеваемости посещающего школу ребенка или уровень бедности домохозяйства, влияют такие факторы, как регион, социально-экономический статус, пол и возраст.

3. Получение ответов на такого рода вопросы требует проведения подробных анализов данных на уровне домохозяйства или отдельного человека. Публикация результатов таких анализов должна, по необходимости включать соответствующие меры точности или правильности оценок, получаемых из данных по результатам обследования. Информация о точности оценок обследо-

вания требуется для надлежащего использования и интерпретации результатов обследования, а также для оценки и совершенствования планов и процедур выборки. Такой контроль и оценка планов выборки особенно важны в случае задействования крупных национальных программ обследований, которые зачастую планируются в качестве единственного источника подробной информации по самой разнообразной тематике.

4. Одной из важнейших единиц измерения точности в выборочных обследованиях выступает вариантность выборки (это понятие рассматривается в главе 3), т. е. индикатор вариантности, появляющейся в результате формирования выборки вместо регистрации данных по всему населению, исходя из того, что собранная в ходе обследования информация достоверна. Вариантность выборки — это мера вариантности распределения оценки по всей выборке. Стандартная ошибка или квадратный корень величины вариантности используется для измерения ошибки выборки. Для любого заданного обследования может быть определена оценка такой ошибки выборки и использована для обозначения достоверности оценок.

5. Форма алгоритма оценки вариантности и способ его определения зависят от используемого в каждом конкретном случае плана выборки. Для стандартных планов выборки такие алгоритмы часто выводятся с помощью простых формул. Однако для используемых в обследованиях домашних хозяйств сложных планов выборки, которые нередко включают в себя стратификацию, формирование кластеров и неравновероятностную выборку, форма таких алгоритмов оценки зачастую отличается сложностью выражения и трудностями при определении. Для расчета ошибок выборки в данном случае необходимы процедуры, учитывающие сложность плана выборки, из которой получены данные, что в свою очередь часто требует использования надлежащего программного обеспечения.

6. Во многих развивающихся странах анализ данных по результатам обследований домашних хозяйств нередко ограничивается базовым анализом таблиц с расчетами средних значений, пропорций и суммарных показателей, при этом не указывается точность или достоверность этих оценок. Даже в национальных статистических службах, имеющих обширную инфраструктуру для сбора и обработки статистических данных, часто можно обнаружить отсутствие опыта подробного анализа данных на микроуровнях. Некоторые разработчики обследований или аналитики зачастую с удивлением узнают, например, что формирование элементов в кластеры вводит корреляцию между элементами, которая снижает точность оценок по сравнению с простой случайной выборкой, которую они привыкли анализировать; или что использование весов в анализе, как правило, увеличивает ошибки выборки; или что стандартные программные пакеты, которые они постоянно используют в работе, надлежащим образом не учитывают такую потерю точности.

7. В этой главе делаются попытки исправить такое положение дел путем предоставления краткого обзора методов расчета величин ошибки выборки для тех видов сложных планов выборки, которые обычно применяются в обследованиях домашних хозяйств в развивающихся странах, а также обзора статистических программных пакетов, используемых для анализа данных таких обследований. Приводится также несколько численных примеров для иллюстрации рассматриваемых процедур определения вариантности.

7.1.2. Обзор

8. В разделе 7.2 дается новый метод определения вариантности выборки применительно к простой случайной выборке, включая численные примеры, иллюстрирующие расчет вариантности выборки и построения доверительных интервалов. Определения других мер оценки ошибок выборки приводятся в разделе 7.3. Раздел 7.4 содержит формулы расчета вариантности выборки при различных стандартных вариантах планов выборки, таких как стратифицированная и гнездовая

выборка. Для облегчения восприятия этих концепций приводится несколько численных примеров. В разделе 7.5 рассматриваются общие характеристики планов выборки обследований домашних хозяйств, а также содержание и структура данных обследования, которые необходимы для надлежащей оценки ошибок выборки. Представлен также общий формат оценок, представляющих основной интерес в обследованиях домашних хозяйств. В разделе 7.6 даются краткие рекомендации касательно представления информации по ошибкам выборки, а в разделе 7.7 дается описание практических методов расчета ошибок выборки при использовании более сложных планов выборки. Эти методы зачастую требуют специальных процедур и использования специализированного программного обеспечения. Ошибки при использовании стандартных статистических компьютерных программ для анализа данных по результатам обследования рассматриваются в разделе 7.8 с помощью примеров, основанных на данных проведенного в 1989 году в Бурунди обследования для оценки охвата населения вакцинацией. В разделах 7.9 и 7.10 дается сравнительный обзор некоторых широко доступных программных пакетов для оценки ошибок выборки в обследованиях домашних хозяйств. В завершающем главу разделе 7.11 приводится ряд заключительных замечаний.

7.2. Вариантность выборки при проведении простой случайной выборки

9. Вариантность той или иной оценки в рамках выборки может быть определена как среднеквадратическое отклонение от среднего значения данной оценки, где такое среднее значение берется из всех возможных вариантов выборки. Как указано в главе 3, простая случайная выборка является наиболее элементарным из всех методов выборки, однако, она редко используется в крупномасштабных обследованиях из-за крайне низкой эффективности в ходе осуществления и чрезмерной дороговизны.

10. Для облегчения понимания концепции вариантности выборки мы рассмотрим небольшую совокупность в составе пяти домохозяйств (N=5), из которых отбирается малая выборка из двух домохозяйств (размер n=2) путем простой случайной выборки без замещения (ПСВБЗ). Предположим, что интересующей нас переменной величиной являются ежемесячные расходы домохозяйства на продукты питания, причем эти расходы каждого из пяти домохозяйств приводятся в таблице 7.1, ниже:

Таблица 7.1
Ежемесячные расходы в долларах на продукты питания в расчете на одно домохозяйство

Домохозяйство (1)	Расходы на продукты питания в долларах (Y_1)
1	10
2	20
3	30
4	40
5	50

11. Прежде всего отметим, что поскольку нам известна интересующая нас переменная величина для всех домохозяйств нашей совокупности, мы можем рассчитать величину параметра, соответствующего среднемесячным расходам домохозяйств на продукты питания, а именно:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{10+20+30+40+50}{5} = \frac{150}{5} = 30.$$

Алгоритм оценки ПСВБЗ для среднемесячных расходов на продукты питания определяется как

$$\hat{\bar{Y}} = \frac{1}{2}\sum_{i\varepsilon S} Y_i,$$

где суммируются показатели по единицам, включенным в выборку. Ясно, что полученная оценка зависит от сформированной выборки. В таблице 7.2, ниже, показаны все возможные выборки; оценки по каждой выборке; отклонения оценки каждой выборки от среднего значения совокупности и квадратичные отклонения. Отметим, что $\hat{\bar{Y}}_{ave}$ указывает среднее значение всех оценок на основе выборок. Отметим также, что $\bar{Y}$ представляет среднее значение совокупности, в то время как $\hat{\bar{Y}}$ дает оценку среднего значения совокупности, называемую выборочным средним значением (см. таблицу 3.2).

Отметим, что среднее значение оценок, основанных на всех возможных выборках, составляет

$$\hat{\bar{Y}}_{ave} = \frac{1}{10}\sum_{i=1}^{10}\hat{\bar{Y}}_i = \frac{15+20+25+30+25+30+35+35+40+45}{10} = \frac{300}{10} = 30 = \bar{Y}.$$

Таблица 7.2
Расчет истинной вариантности выборки касательно значения $\hat{\bar{Y}}$, параметра для определения среднего значения

Выборка 1	Единицы выборки	Оценка для выборки ($\hat{\bar{Y}}_i$)	$\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave}$	$(\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave})^2$
1	(1, 2)	15	-15	225
2	(1, 3)	20	-10	100
3	(1, 4)	25	-5	25
4	(1, 5)	30	0	0
5	(2, 3)	25	-5	25
6	(2, 4)	30	0	0
7	(2, 5)	35	5	25
8	(3, 4)	35	5	25
9	(3, 5)	40	10	100
10	(4, 5)	45	15	225
Среднее значение		**30**	**0**	**750**

12. Иными словами, среднее значение оценки по всем возможным выборкам равно среднему значению совокупности. Оценка с такими характеристиками называется *несмещенной* в отношении определяемого параметра.

13. Истинная вариантность выборки в отношении оцениваемых среднемесячных расходов на продукты питания на основе ПСВБЗ размером n=2 из данной совокупности составляет

$$Var(\hat{\bar{Y}}) = \frac{1}{10}\sum_{i=1}^{10}(\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave})^2 = \frac{750}{10} = 75.$$

14. Проблема при использовании приведенного выше подхода обусловлена тем фактом, что с практической точки зрения нецелесообразно формировать все возможные выборки из данной совокупности. На практике создается только одна выборка, при этом значения совокупности

неизвестны. Более применимым в реальности подходом является использование формул для расчета вариантности. Такие формулы существуют для всех стандартных планов выборки.

15. При простой случайной выборке без замещения, вариантность выборки для оцениваемого среднего значения ($\hat{\bar{Y}}$), основанная на размере выборки, n, определяется выражением

$$Var(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{\delta^2}{n}. \qquad (7.1)$$

где $\delta^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{\bar{Y}})^2}{N-1}$ —

это мера вариантности интересующей характеристики (вариантность совокупности Y). Обычно, величина δ^2 неизвестна и должна быть оценена исходя из выборки (см. уравнение 7.2 ниже). Из приведенной выше формулы ясно видно, что вариантность выборки зависит от следующих факторов:

a) вариантности совокупности по интересующей характеристике;

b) размера совокупности;

c) размера выборки;

d) плана выборки и метода оценки.

16. Доля включенной в выборку совокупности, n/N, называется долей выборки (обозначается символом f); а множитель $[1 - (n/N)]$ или $1–f$, который является долей совокупности, не включенной в выборку, называется поправкой для конечной совокупности (*fpc*). *fpc* представляет собой поправку, вносимую в стандартную ошибку оценки для учета того факта, что выборка формируется без замещения из конечной совокупности. Отметим, однако, что при небольшой доле выборки величину *fpc* можно не учитывать. На практике величину *fpc* можно не учитывать, если она не превышает 5 процентов (Cochran, 1977).

17. Приведенная выше формула указывает, что вариантность выборки обратно пропорциональна размеру выборки. По мере увеличения размера выборки вариантность выборки снижается; причем применительно к переписи или поголовной регистрации (где n=N), вариантности выборки вообще не существует. Отметим, что неполучение ответов реально сокращает размер выборки и, таким образом, повышает вариантность выборки.

18. Можно показать, что несмещенная оценка вариантности выборки в отношении оцениваемого среднего значения выражается как

$$v(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{s^2}{n} \qquad (7.2)$$

где $s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\bar{Y}})^2}{n-1}$ —

это оценка вариантности совокупности, δ^2, основанная на данной выборке. Она называется вариантностью выборки. 95-процентный доверительный интервал для среднего значения совокупности (см. пункт 30 главы 3) выражается как

$$\hat{\bar{Y}} \pm 1{,}96\sqrt{v(\hat{\bar{Y}})}. \qquad (7.3)$$

19. Для доли совокупности основанная на выборке оценка и оцениваемая вариантность выражаются, соответственно, как

$$\hat{P} = \frac{\textit{число единиц, обладающих искомой характеристикой}}{n} \qquad (7.4)$$

и $$v(\hat{P}) = (1 - \frac{n}{N})\frac{\hat{P}(1-\hat{P})}{n-1}. \qquad (7.5)$$

20. В таблице 7.3, ниже, дается резюме оценок различных параметров совокупности и уровней вариантности этих оценок при простой случайной выборке без замещения.

Таблица 7.3
Оценки и их уровни вариантности для отдельных характеристик совокупности

Параметр	Оценка	Вариантность оценки
Среднее значение совокупности ($\hat{\bar{Y}}$)	$\hat{\bar{Y}} = \frac{1}{n_{\text{яс}}} \sum_{Sample} Y_i$	$v(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{s^2}{n}$
Общее значение совокупности	$\hat{T} = N\hat{\bar{Y}}$	$v(\hat{T}) = N^2 v(\hat{\bar{Y}})$
Доля совокупности для той или иной категории	$\hat{P} = \frac{\textit{число единиц в данной категории, вошедших в выборку}}{n}$	$v(\hat{P}) = (1 - \frac{n}{N})\frac{\hat{P}(1-\hat{P})}{n-1}$

21. Как правило, $(1-\alpha)$-процентный доверительный интервал для среднего значения совокупности выражается как

$$\textit{Оценка} \pm z_{1-\alpha/2}\sqrt{\textit{расчетная вариантность оценки}} \qquad (7.6)$$

где $z_{1-\alpha/2}$ — это $(1-\alpha/2)$-я процентиль стандартного нормального распределения.

22. Приведенный ниже пример иллюстрирует оценку вариантности выборки, основанную на сформированной выборке.

Пример 1

Рассмотрим простую случайную выборку из n = 20 домохозяйств, отобранную из крупной совокупности N=20 000 домохозяйств. Собранные данные представлены в таблице 7.4, ниже, где переменная величина Y обозначает ежемесячные расходы на продукты питания, а переменная величина Z обозначает наличие или отсутствие телевизора в домохозяйстве (z=1, если да, и z=0, если нет).

Таблица 7.4

Еженедельные расходы домохозяйств на продукты питания и владение телевизором для вошедших в выборку домохозяйств

Домохозяйство (i)	Y_i	Z_i	i	Y_i	Z_i
1	5	0	11	7	1
2	10	1	12	8	1
3	5	0	13	9	1
4	9	1	14	10	1
5	5	1	15	8	1
6	6	1	16	8	0
7	7	0	17	5	0
8	15	1	18	7	0
9	12	1	19	12	1
10	8	0	20	4	0
Всего		160	12		

Оценка среднего значения по совокупности для ежемесячных расходов домохозяйства на продукты питания составляет

$$\hat{\bar{Y}} = \frac{1}{20}\sum_{i=1}^{20} Y_i = \frac{5+10+\ldots+12+4}{20} = \frac{160}{20} = 8.$$

Расчетная величина вариантности оцениваемого среднего значения составляет

$$v(\hat{\bar{Y}}) = (1-\frac{20}{20\ 000})\left\{\frac{(5-8)^2+(10-8)^2+\ldots..+(12-8)^2+(4-8)^2}{19}\right\} = 7{,}87.$$

95-процентный доверительный интервал для среднего значения совокупности составляет

$$8 \pm 1{,}96 \times \sqrt{7{,}87} = (2{,}50,\ 13{,}50).$$

Оценка общих по совокупности расходов домохозяйств на продукты питания составляет

$$\hat{Y} = N\hat{\bar{Y}} = 20\ 000 \times 8 = 160\ 000.$$

Расчетная величина вариантности оцениваемого общего значения составляет

$$v(\hat{Y}) = 20\ 000^2 \times 7{,}87 = 3\ 148\ 000\ 000.$$

95-процентный доверительный интервал для среднего значения совокупности составляет

$$160\ 000 \pm 1{,}96 \times \sqrt{3\ 148\ 000\ 000} = (50\,030,\ 269\,970)\ .$$

Оценка доли в совокупности домохозяйств, имеющих телевизоры, составляет

$$\hat{P} = \frac{1}{20}\sum_{i=1}^{20} Z_i = \frac{12}{20} = 0{,}6.$$

Расчетная величина вариантности оцениваемой доли домохозяйств с телевизорами составляет

$$v(\hat{P}) = (1-\frac{20}{20\ 000})\frac{0{,}6(1-0{,}6)}{19} = 0{,}0126.$$

95-процентный доверительный интервал для среднего значения совокупности составляет

$$0{,}6 \pm 1{,}96 \times \sqrt{0{,}0126} = (0{,}38,\ 0{,}82).$$

7.3. Прочие виды измерения ошибок выборки

23. В дополнение к вариантности выборки существуют и другие измерители ошибок выборки. К ним относятся стандартная ошибка, коэффициент вариации и эффект схемы. Все эти измерители алгебраически связаны между собой в том смысле, что выражение любого из этих измерителей можно вывести из других с помощью простых алгебраических действий.

7.3.1. Стандартная ошибка

24. Стандартная ошибка алгоритма оценки — это квадратный корень вариантности его выборки. Этот измеритель легче интерпретировать, поскольку он предоставляет величину ошибки выборки, используя такой же масштаб, что и для оценки, в то время как вариантность базируется на квадратах разностей.

25. Вопрос, часто возникающий при составлении плана обследования, состоит в том, сколь высокая стандартная ошибка может считаться приемлемой. Ответ на этот вопрос зависит от порядка величины оценки. Например, стандартная ошибка, равная 100, будет считаться небольшой при оценке годового дохода, но большой при оценке среднего веса людей. Таким образом, стандартная ошибка величиной $\sqrt{3\ 148\ 000\ 000} = 56\ 107$ для расчетной суммарной величины расходов в 160 000, полученной выше в примере 1, может считаться слишком большой.

7.3.2. Коэффициент вариации

26. Коэффициент вариации (КВ) оценки — это отношение стандартной ошибки к среднему значению самой оценки. Следовательно, КВ позволяет измерить ошибку выборки относительно измеряемой характеристики. Обычно этот коэффициент выражается в процентах.

27. КВ полезно применять при сравнении точности оценок обследования, имеющих различные размеры или масштабы. Однако он неприменим для оценки характеристик, истинное значение которой может равняться нулю или быть отрицательной, включая оценки изменения, например изменения среднего дохода за двухлетний период.

7.3.3. Эффект схемы

28. Эффект схемы (обозначаемый как *deff*) определяется как отношение вариантности выборки той или иной оценки при данном плане выборки к вариантности выборки данной оценки, основанной на простой случайной выборке такого же размера. Он может рассматриваться как коэффициент, на который необходимо умножить вариантность оценки, основанной на простой случайной выборке такого же размера, для учета сложностей фактического плана выборки в связи с такими факторами, как стратификация, формирование кластеров и взвешивание.

29. Иными словами, оценка, базирующаяся на данных из сложной выборки размера n, имеет такую же вариантность, что и оценка, рассчитанная по данным из простой случайной выборки размера $n/deff$. По этой причине отношение $n/deff$ иногда называют эффективным размером выборки для оценки, базирующейся на данных из сложного плана выборки. Для общего рассмотрения вопроса расчетов «эффективного размера выборки» см. работы Киша (1995 год) и Поттхоффа, Вудбери и Мантона (1992 год) и указанные в них ссылки. Кроме того, более подробное рассмотрение эффектов схемы и их использования в плане выборки см. в различных разделах главы 3.

7.4. Расчет вариантности выборки для других стандартных планов

30. Для простых планов выборки и простых линейных оценок, таких как средние значения, пропорции и суммарные величины, обычно существует возможность вывести формулы, которые можно использовать для расчета вариантности оценок. Однако для тех видов сложных планов выборки и оценок, которые обычно ассоциируются с обследованиями домашних хозяйств, это зачастую сделать затруднительно или невозможно. В этом разделе мы приводим примеры для иллюстрации расчета вариантности выборки для стратифицированных планов и одноэтапных планов гнездовой выборки. Формулы и примеры расчетов вариантности для других стандартных планов выборки даются в учебниках (например, Кокран, 1977 год; и Киш, 1965 год).

7.4.1. Стратифицированная выборка

31. Подробное описание стратифицированной выборки приводится в главе 3 и приложении I. В данном разделе мы затронем лишь оценку вариантности для данного плана выборки. Рассмотрим стратифицированную выборку с H стратами и с выборочными оценками средних значений совокупности для страт, выраженными как $\bar{Y}_1, \bar{Y}_2, \ldots \ldots \bar{Y}_H$, а также с выборочными оценками вариантности совокупности для страт, выраженными как $S_1^2, S_2^2, \ldots \ldots S_H^2$. Алгоритмом оценки среднего значения совокупности для данного плана выборки является уравнение

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{H} \hat{\bar{Y}}_h, \tag{7.7}$$

где $\hat{\bar{Y}}_h$ — это основанная на выборке оценка $\bar{Y}_h$, $h = 1, \ldots \ldots H$. Вариантность оценки определяется как:

$$v(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{H} v(\hat{\bar{Y}}_h). \tag{7.8}$$

При стратифицированной случайной выборке алгоритм оценки и его расчетная вариантность определяется как:

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} W_h \bar{y}_h, \tag{7.9}$$

где $\bar{y}_h$ — это выборочное среднее значение для страты h, N_h — это размер совокупности в страте h, а

$$W_h = \frac{N_h}{N}, \; h = 1, \ldots \ldots H.$$

Расчетная вариантность данной оценки при стратифицированной случайной выборке выводится по формуле:

$$v(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{H} W_h^2 v(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 (1 - \frac{n_h}{N_h}) \frac{s_h^2}{n_h}, \tag{7.10}$$

где n_h — это размер выборки в страте h, а S_h^2 — вариантность выборки, основанная на выборке оценка S_h^2, h=1, …, H.

Пример 2

Теперь применим эти результаты в отношении примера стратифицированного плана выборки, предусматривающего три страты с параметрами, которые приводятся в таблице 7.5, ниже. Предположим, что нас интересует оценка среднего значения совокупности, основанного на общем размере выборки в 1 500 единиц.

Таблица 7.5
Примеры данных для стратифицированного плана выборки

Параметр	Совокупность	Страта 1 (столица)	Страта 2 (провинция — городские районы)	Страта 3 (провинция — городские районы)
Размер	N =1 000 000	N_1 = 300 000	N_2 = 500 000	N_3 = 200 000
Вариантность	S^2=75 000	S_1^2 = ?	S_2^2 = ?	S_3^2 = ?
Среднее значение	$\bar{Y}$ = ?	$\bar{Y}_1$ = ?	$\bar{Y}_2$ = ?	$\bar{Y}_3$ = ?
Расходы на единицу	N/A	C_1=1	C_2=4	C_3=16
Размер выборки при оптимальном распределении[a]	n=1 500	n_1=857	n_2=595	n_3=48
Выборочное среднее	N/A	$\bar{y}_1 = 4\ 000$	$\bar{y}_2 = 2\ 500$	$\bar{y}_3 = 1\ 000$
Вариантность выборки	N/A	$s_1^2 = 90\ 000$	$s_2^2 = 62\ 500$	$s_3^2 = 10\ 000$

Примечание: N/A — неприменимо.
[a] См. главу 3.

Оценка среднего значения совокупности равна

$$\hat{\bar{Y}}_{st} = \frac{300\ 000}{1\ 000\ 000} \times 4\ 000 + \frac{500\ 000}{1\ 000\ 000} \times 2\ 500 + \frac{200\ 000}{1\ 000\ 000} \times 1\ 000 = 2\ 650.$$

Расчетная вариантность указанной выше оценки равна

$$v(\hat{\bar{Y}}_{st}) = \left(\frac{300\ 000}{1\ 000\ 000}\right)^2 \left(1 - \frac{857}{300\ 000}\right)\left(\frac{90\ 000}{857}\right) + \left(\frac{500\ 000}{1\ 000\ 000}\right)^2 \left(1 - \frac{595}{500\ 000}\right)\left(\frac{62\ 500}{595}\right) +$$
$$\left(\frac{200\ 000}{1\ 000\ 000}\right)^2 \left(1 - \frac{48}{200\ 000}\right)\left(\frac{10\ 000}{48}\right) = 43{,}98516.$$

95-процентный доверительный интервал для среднего значения совокупности составляет

$$2\ 650 \pm 1{,}96 \times \sqrt{43{,}98516} = (2\ 637,\ 2\ 663).$$

Отметим, что расчетная вариантность оценки среднего значения при простой случайной выборке выражается уравнением:

$$v(\hat{\bar{Y}}_{SRS}) = \left(1 - \frac{1500}{1\ 000\ 000}\right) \times \frac{75\ 000}{1500} = 49{,}925.$$

Таким образом, эффект схемы этого стратифицированного плана выборки равен $\frac{43{,}98516}{49{,}925} = 0{,}88$, а эффективный размер выборки равен $\frac{1\,500}{0{,}88} = 1705$.

Это означает, что оценка, основанная на стратифицированной случайной выборке размером в 1500 единиц, имеет такую же вариантность, что и оценка, основанная на простой случайной выборке размером в 1705 единиц.

32. Подробное описание метода гнездовой выборки приводится в главе 3. В данном разделе мы даем простой пример для иллюстрации расчета ошибок выборки для особого случая одноэтапной гнездовой выборки.

Пример 3

Предположим, что нас интересует оценка доли детей школьного возраста, проживающих в той или иной провинции, которые прошли вакцинацию от полиомиелита. Для простоты допустим, что в общей сложности имеются 500 счетных участков (СУ) равного размера, в каждом из которых проживает по 25 детей школьного возраста. Счетные участки в данном примере служат в качестве кластеров. Предположим, что из 500 счетных участков в данной провинции мы отберем 10 СУ путем простой случайной выборки без замещения и что доля вакцинированных детей школьного возраста измерялась в каждом вошедшем в выборку СУ, а результаты этого измерения представлены в таблице 7.6, ниже.

Таблица 7.6
Доля прошедших вакцинацию детей школьного возраста в 10 счетных участках как искомая переменная величина

Вошедший в выборку СУ (i)	1	2	3	4	5	6	7	8	9	10
Доля выборки $\hat{P}_i$	$\frac{8}{25}$	$\frac{10}{25}$	$\frac{12}{25}$	$\frac{14}{25}$	$\frac{15}{25}$	$\frac{17}{25}$	$\frac{20}{25}$	$\frac{20}{25}$	$\frac{21}{25}$	$\frac{23}{25}$

Для этого примера оценка доли вакцинированных детей школьного возраста в данной провинции составляет

$$\hat{P} = \frac{160}{250} = 0{,}64 \text{ или } 64\%.$$

Далее, вариантность выборки составляет

$$s_{\hat{P}}^2 = \frac{1}{10-1}\sum_{i=1}^{10}(\hat{P}_i - \hat{P})^2 = 0{,}040533.$$

Таким образом, вариантность оцениваемой доли составляет

$$v(\hat{P}) = (1-\frac{10}{500})\times\frac{0{,}040533}{10} = \ 0{,}003972.$$

Отметим, что при простой случайной выборке расчетная вариантность оцениваемой доли составляет

$$v(\hat{P}_{SRS}) = (1-\frac{250}{12\,500})\times\frac{0{,}64(1-0{,}64)}{250-1} = \ 0{,}0009078.$$

Таким образом, эффект схемы для данного плана гнездовой выборки составляет $\frac{0{,}003972}{0{,}0009078} = 4{,}38$, а эффективный размер выборки равен $\frac{250}{4{,}38} = 57$.

Это означает, что оценка, основанная на гнездовой выборке размером в 250 единиц, имеет такую же вариантность, что и оценка, основанная на простой случайной выборке размером в 57 единиц.

7.5. Общие характеристики планов выборки при обследованиях домашних хозяйств и данных по результатам обследований

7.5.1. Отклонения планов выборки при обследованиях домашних хозяйств от простой случайной выборки

33. Как указывалось выше, на практике простая случайная выборка редко используется в крупномасштабных обследованиях домашних хозяйств из-за слишком высоких расходов на ее осуществление. Тем не менее важно досконально разбираться в этом плане, поскольку он является теоретической основой для более сложных планов выборки. Большинство планов выборки для обследований домашних хозяйств отличаются от простой случайной выборки присутствием одной или нескольких следующих трех характеристик:

a) стратификация на одном или нескольких этапах выборки;

b) объединение единиц в кластеры на одном или нескольких этапах выборки, что снижает расходы, но повышает вариантность оценок вследствие корреляции между единицами внутри одного и того же кластера;

c) взвешивание для компенсации таких дефектов выборки, как неравная вероятность отбора, неполучение ответов и неполный охват (более подробную информацию см. в главе 6).

34. План выборки считается *сложным*, если он имеет одну или несколько из приведенных выше характеристик. Большинство планов выборки при обследованиях домашних хозяйств являются сложными и, следовательно, не удовлетворяют допущениям простой случайной выборки. Вследствие этого проведение анализа данных по результатам обследования домашних хозяйств таким же образом, как если бы они были получены исходя из плана простой случайной выборки, приведет к ошибкам в анализе и в тех выводах, которые делаются на основе таких данных. Более того, как уже упоминалось, представляющие интерес оценки большинства обследований домашних хозяйств не могут быть выражены в виде линейных функций наблюдений, и поэтому могут отсутствовать формулы в закрытой форме для расчета уровней вариантности. В следующих разделах рассматривается вопрос о методах оценки вариантности для различных планов выборки при обследованиях домашних хозяйств, в которых принимаются во внимание упомянутые выше усложняющие факторы.

7.5.2. Подготовка файлов данных для проведения анализа

35. Собираемые в развивающихся странах данные по результатам обследований подчас не поддаются анализу, выходящему за рамки базисных плотностей распределения и составления таблиц. Для этого существует несколько причин. Во-первых, может быть в наличии весьма ограниченная или вообще отсутствовать техническая документация по плану выборки обследования. Во-вторых, файлы данных могут не иметь того формата, структуры и не содержать требуемой информации, что позволило бы провести какой-либо сложный анализ. В-третьих, могут отсутствовать надлежащее программное обеспечение и технический опыт.

36. Для проведения надлежащего анализа данных по результатам выборочного обследования домашних хозяйств связанная с ним база данных должна содержать всю информацию, отражающую процесс формирования выборки (подробное обсуждение этого вопроса см. в главе 5). В частности, база данных должна включать соответствующие маркировки для предусмотренных планом выборки страт, первичных единиц выборки (ПЕВ), вторичных единиц выборки (ВЕВ) и т. д. Иногда в целях расчета уровней вариантности возникает необходимость внесения изменений в страты и ПЕВ, которые фактически используются в формировании выборки. Такие изменения необходимы для того, чтобы фактический план выборки соответствовал варианту плана выборки, включенному хотя бы в один из программных пакетов для статистического анализа (см. раздел 7.9). Создаваемые для оценки вариантности страты и ПЕВ иногда называют псевдостратами или стратами вариантности и псевдо-ПЕВ или ПЕВ вариантности. Необходимые переменные плана выборки, а также переменные, создаваемые в целях оценки вариантности, должны включаться в файл данных вместе с соответствующей документацией о том, как определяются и используются эти переменные. Для расчета вариантности требуется минимальный набор из трех переменных: вес выборки, страта (или псевдострата) и ПЕВ (или псевдо-ПЕВ). Эти три переменных завершают создание плана выборки, и их включение в данные по результатам обследования позволяет проводить надлежащий анализ этих данных с учетом усложняющих факторов в плане выборки.

37. Более того, для каждой единицы выборки в файле данных должны быть разработаны веса выборки. Эти веса должны отражать вероятность отбора каждой единицы выборки, а также компенсировать случаи неполучения ответов и другие дефекты выборки. Веса выборки и маркировки переменных плана выборки требуются для надлежащего расчета вариантности оценок обследования. Как упоминалось в главе 6 и в предыдущих разделах этой главы, веса выборки важны не только для получения соответствующих оценок обследования, но и для определения ошибок выборки для этих оценок. Таким образом, чрезвычайно важно, чтобы вся информация по весам была включена в файлы данных. В частности, при внесении любых поправок на неполучение ответов, на последующую стратификацию или иных корректировок документация по обследованию должна содержать описание процедур внесения таких поправок.

7.5.3. Виды оценок обследования

38. Для большинства обследований домашних хозяйств наиболее распространенные и вызывающие интерес оценки выражаются в форме суммарных величин и соотношений. Представим себе стратифицированный трехэтапный план выборки, состоящий из ПЕВ на первом этапе, ВЕВ на втором этапе и домохозяйств на третьем этапе. Оценка обследования может быть выражена следующим образом:

$$\hat{Y}=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_i}\sum_{k=1}^{l_j} W_{hijk}\, Y_{hijk}, \qquad (7.11)$$

где W_{hijk} = итоговый вес k-го домохозяйства (k = 1, … … 1_j), отобранного в j-ой ВЕВ (j = 1, … … m_i) и в i-ой ПЕВ (i = 1, … … n_h) в h-ой страте (h = 1, … … H); а

Y_{hijk} = величина переменной Y для k-го домохозяйства, отобранного в j-ой ВЕВ и в i-ой ПЕВ в h-ой страте.

39. На наибольшем базисном уровне веса, связанные с единицами выборки, обратно пропорциональны вероятностям отбора данных единиц в выборку. Однако зачастую используются более

сложные методы для вычисления весов, которые будут применяться в ходе анализа данных. Некоторые из этих методов описаны в главе 6 и в приведенных в ней ссылках.

40. Оценка выборки, выраженная в форме соотношения, равна

$$\hat{R} = \frac{\hat{Y}}{\hat{X}}, \tag{7.12}$$

где $\hat{Y}$ и $\hat{X}$ — это оценки суммарных величин переменных Y и X, соответственно, рассчитанных по уравнению 7.12, выше.

41. В случае многоэтапной выборки средние значения и пропорции представляют собой лишь особые случаи оценок соотношений. Применительно к среднему значению, переменная X в знаменателе этого соотношения является числовой переменной и определяется как равная 1 для всех элементов с тем, чтобы знаменатель представлял собой сумму весов. Применительно к той или иной пропорции, переменная X в знаменателе также определяется как равная 1 для всех элементов; а переменная Y в числителе — это биноминальная переменная, определяемая как равная либо 0, либо 1, в зависимости от наличия или отсутствия в наблюдаемой единице той характеристики, доля которой рассчитывается. В большинстве обследований домашних хозяйств знаменателем в алгоритме расчета соотношения является общая численность населения, общая численность мужчин, общая численность женщин, общая численность сельского населения, общая численность населения данной провинции или района и т. д.

7.6. Рекомендации по представлению информации об ошибках выборки

7.6.1. Определение областей отчетности

42. В крупномасштабных обследованиях домашних хозяйств с многочисленными исследуемыми переменными величинами и областями выборки и с несколькими, подчас конкурирующими друг с другом целями, практически невозможно представлять каждую отдельную оценку вместе с относящейся к ней ошибкой выборки. Это не только значительно увеличит объем публикации, но и, скорее всего, создаст помехи в представлении результатов по существу обследования. С учетом ожидаемой вариантности самих оценок ошибки выборки, представление результатов по слишком многим отдельным переменным величинам может привести к путанице и ощущению нестабильности качества всех собранных данных в ходе обследования. Гораздо более целесообразно представлять информацию по ошибкам выборки для нескольких наиболее важных и представляющих интерес характеристик, помещая остальные данные в добавление.

43. При представлении информации об ошибках выборки важно учитывать ее возможное влияние на толкование результатов обследования и политические решения, которые могут быть приняты на основе такого толкования. Информация об ошибках выборки должна всегда рассматриваться как лишь один из элементов общей ошибки выборки в ходе обследования, и далеко не всегда самый существенный. В определенных ситуациях, складывающихся при проведении обследования, ошибки регистрации данных (см. главу 8) могут оказывать более значительное влияние на общее качество данных по результатам обследования, чем ошибки выборки. По этой причине рекомендуется включать в информацию об ошибках выборки описание основных источников ошибок регистрации данных и некоторые качественные оценки их влияния на общее качество данных по результатам обследования. Поскольку влияние ошибок выборки резко возрастает на более низких уровнях раз-

укрупнения единиц, рекомендуется также включать некоторые предупреждающие замечания по поводу допустимой степени разукрупнения данных, полученных в ходе обследования.

44. Как правило, информация об ошибках выборки должна быть достаточно подробной, чтобы облегчить правильное толкование результатов обследования и удовлетворить требования всего спектра пользователей данных — от обычного пользователя или лица, принимающего решения (заинтересованного в использовании результатов обследования для формулирования политики), или специалиста, занимающегося предметным анализом (осуществляющего дальнейший анализ результатов и составляющего отчеты по ним), до статистика, формирующего выборку (который решает проблему статистической эффективности данного плана выборки в сравнении с другими вариантами и с теми или иными характеристиками данного плана, которые можно будет использовать при разработке планов будущих обследований).

7.6.2. Как составить отчет об ошибках выборки

45. Ошибки выборки могут быть представлены в трех различных формах, а именно, как:

a) абсолютные значения стандартных ошибок;

b) относительные стандартные ошибки (квадратные корни относительных значений вариантности);

c) доверительные интервалы.

46. Выбор одной из этих трех форм представления данных зависит от характера оценки. В случаях, когда оценки варьируются по размерам и по единицам измерения, одни и те же значения стандартных ошибок могут применяться к оценкам, если они выражены в относительной форме; следовательно, более эффективно указывать относительные стандартные ошибки. При этом, однако, абсолютные значения стандартных ошибок намного легче понимать и относить к оценкам, особенно в случаях процентных долей, пропорций и коэффициентов. Использование доверительных интервалов требует выбора доверительного уровня (скажем, 90, 95 вмененных или 99 процентов). Поскольку они варьируются в зависимости от целей обследования и требований к точности оценок, важно указать доверительный уровень, используемый при представлении информации об ошибках выборки, а затем сохранять данный доверительный уровень на всем протяжении обследования при определении значимости полученных результатов. Как указывалось выше, наиболее часто на практике используется 95-процентный доверительный интервал (см. пункты 30 и 22 в главах 3 и 7, соответственно), а именно:

$$\text{Оценка} \pm 1{,}96 \times \text{Стандартная ошибка} \tag{7.13}$$

47. Более подробное обсуждение вопроса о представлении информации по ошибкам выборки, включая конкретные рекомендации для различных категорий пользователей и целый ряд примеров, см. в публикации Организации Объединенных Наций (1993 год) и в указанных в ней ссылках.

7.6.3. Эмпирическое правило составления отчетов о стандартных ошибках

48. Часто применяемое в отчетах о стандартных ошибках эмпирическое правило заключается в том, что стандартная ошибка указывается в виде двух значащих разрядов, а затем приводится соответствующая точечная оценка с таким же числом десятичных разрядов, что и в стандартной ошибке. Например:

1. Если точечная оценка составляет 73 456 при стандартной ошибке в 2 345, тогда в отчете мы указываем точечную оценку как 73 500 и стандартную ошибку как 2 300.
2. Если точечная оценка составляет 1,54328 при стандартной ошибке в 0,01356, тогда в отчете мы указываем точечную оценку как 1,543 и стандартную ошибку как 0,014.

49. Общее обоснование этого эмпирического правила, по-видимому, связано с t-статистикой. Присутствие двух значимых разрядов в стандартной ошибке и соответствующее число разрядов в точечной оценке гарантирует не слишком большое влияние ошибки округления на результирующий показатель t-статистики и, в то же время, устраняет эффект излишнего уровня точности, заданного точечными оценками с большим числом несущественных разрядов. Отметим, однако, что это эмпирическое правило не всегда работает в условиях, когда t-статистика не представляет особого интереса.

7.7. Методы расчета вариантности при обследованиях домашних хозяйств

50. В данном разделе мы приводим краткое описание некоторых обычных методов расчета вариантности или ошибок выборки для оценок, базирующихся на данных по результатам обследования. Методы расчета ошибок выборки для обследований домашних хозяйств можно разделить на четыре больших категории:

a) точные методы;

b) метод кластера последнего этапа отбора;

c) аппроксимации линеаризации;

d) методы репликации.

Ниже мы кратко рассмотрим каждую из этих категорий по очереди. Заинтересованные читатели могут получить более подробную информацию из таких источников, как Киш и Франкель (1974 год), Уолтер (1985 год), а также Лехтонен и Пахкинен (1995 год).

7.7.1. Точные методы

51. В разделах 7.2 и 7.4 приводится ряд примеров точных методов расчета вариантности для стандартных планов выборки. Эти методы, если они применимы, представляют собой наилучший подход к расчету вариантности. При этом, однако, их применение к расчету вариантности оценок выборки, базирующихся на данных по результатам обследования домашних хозяйств, осложняется несколькими факторами. Во-первых, планы выборки, используемые в большинстве обследований домашних хозяйств, являются более сложными, чем простая случайная выборка (см. раздел 7.5.1, выше). Во-вторых, искомые оценки могут не иметь форму простых линейных функций наблюдаемых величин, и поэтому довольно часто вариантность выборки может не выражаться формулой закрытого типа, как формула выборочного среднего значения при простой случайной выборке или при стратифицированной выборке. Более того, точные методы зависят от плана выборки, рассматриваемого для конкретного случая; искомой оценки и используемых процедур взвешивания.

52. В нижеследующих разделах мы обсудим методы расчета вариантности планов выборки, которые обычно используются для обследований домашних хозяйств. Эти методы предназначены для преодоления недостатков точных методов.

7.7.2. Метод кластера последнего этапа отбора

53. Метод кластера последнего этапа отбора (Hansen, Hurwitz and Madow, 1953, pp. 257–259) может применяться для расчета вариантности оценок в ходе обследования, базирующихся на выборке, которая получена с помощью сложного плана выборки. Согласно этому методу, кластер последнего этапа выборки содержит полную выборку из ПЕВ вне зависимости от формирования выборки на последующих этапах многоэтапного плана. Оценки вариантности рассчитываются, используя исключительно суммарные значения отдельных ПЕВ без необходимости расчета элементов вариантности на каждом этапе отбора.

54. Предположим, что выборка из n_h ПЕВ отбирается из страты h (с любым числом этапов в пределах всех ПЕВ). Тогда оценка суммарного значения для страты h определяется как

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}, \qquad (7.14)$$

где $\hat{Y}_{hi} = \sum_{j=1}^{m_i} W_{hijk}\, Y_{hijk}$.

Отметим, что оценкой $\hat{Y}_{hi}$ на уровне ПЕВ является оценка $\frac{\hat{Y}_h}{n_h}$. Следовательно, вариантность отдельных оценок уровня ПЕВ определяется как

$$v(\hat{Y}_{hi}) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2. \qquad (7.15)$$

а вариантность их суммарного значения, $\hat{Y}_h$, т. е. суммы на уровне страты, рассчитанной из оценки суммарного значения по совокупности для страты h, в рамках случайной выборки размером n_h, определяется как

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2. \qquad (7.16)$$

55. Отметим, что простое алгебраическое действие дает следующее эквивалентное уравнение оценки вариантности для суммарного значения по совокупности для страты h:

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1} \left\{ \sum_{i=1}^{n_h} \hat{Y}_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} \hat{Y}_{hi} \right)^2}{n_h} \right\}. \qquad (7.17)$$

56. И наконец, с помощью независимой выборки между стратами оценка вариантности для суммарного значения по генеральной совокупности получается путем подсчета суммы вариантностей суммарных значений на уровне страт, иными словами,

$$v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h). \tag{7.18}$$

Отметим, что иногда в указанных выше формулах используется коэффициент поправки на конечность совокупности $(1-n_h/N_h)$.

57. Уравнение (7.18) примечательно с той точки зрения, что вариантность оцениваемого суммарного значения является функцией только лишь надлежащим образом взвешенных суммарных значений $\hat{Y}_{hi}$ ПЕВ без какой-либо ссылки на структуру и способ взвешивания в пределах ПЕВ. Это значительно упрощает формулу расчета вариантности, поскольку не требуется вычисления элементов вариантности, относящихся к другим этапам формирования выборки в рамках ПЕВ. Эта особенность обусловливает для метода кластера последнего этапа отбора значительную гибкость при обработке различных планов выборки, и это, по сути, является одной из сильных сторон этого метода и основной причиной его широкого использования в обследованиях.

58. Итак, оценка вариантности отношения $\hat{R} = \dfrac{\hat{Y}}{\hat{X}}$, определяется как

$$v(\hat{R}) = \frac{1}{\hat{X}^2}\left\{v(\hat{Y}) + \hat{R}^2 v(\hat{X}) - 2\hat{R}\,\mathrm{cov}(\hat{Y},\ \hat{X})\right\}, \tag{7.19}$$

где $v(\hat{Y})$и $v(\hat{X})$ рассчитываются по формуле вариантности оцениваемого суммарного значения, и

$$\mathrm{cov}\,(\hat{Y},\hat{X}) = \sum_{h=1}^{H}\left\{\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h}\right)\left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h}\right)\right\}, \tag{7.20}$$

или, что равнозначно,

$$\mathrm{cov}\,(\hat{Y},\hat{X}) = \frac{n_h}{n_h-1}\left\{\sum_{i=1}^{n_h}\hat{X}_{hi}\,\hat{Y}_{hi} - \frac{\left(\sum_{i=1}^{n_h}\hat{X}_{hi}\right)\left(\sum_{i=1}^{n_h}\hat{Y}_{hi}\right)}{n_h}\right\}. \tag{7.21}$$

59. Отметим, что указанная выше формула вариантности отношения может быть упрощена путем использования того факта, что относительная вариантность данного отношения примерно равна разности между значениями относительной вариантности числителя и знаменателя. Напомним, что относительная вариантность оценки — это отношение величины вариантности к ее квадрату. Следовательно, оцениваемое отношение $\hat{R}$ относительной вариантности, обозначаемое через *relvar* $(\hat{R})$, определяется как

$$relvar\,(\hat{R}) = \frac{v(\hat{R})}{\hat{R}^2}\ . \tag{7.22}$$

Таким образом, оценка вариантности отношения определяется как

$$v\,(\hat{R}) = \hat{R}^2 relvar\,(\hat{R}) = \hat{R}^2\left\{relvar\,(\hat{Y}) - relvar\,(\hat{X})\right\}. \tag{7.23}$$

60. Метод кластера последнего этапа отбора для расчета ошибок выборки для оцениваемых суммарных значений и отношений можно схематически описать в виде следующих шагов:

- *Шаг 1.* Для каждой страты по отдельности рассчитываем взвешенную оценку $\hat{Y}_{hi}$ для интересующей характеристики Y по каждой ПЕВ (в соответствии с процедурами взвешивания, указанными в главе 6).
- *Шаг 2.* Рассчитываем квадрат каждой оцениваемой величины ПЕВ, полученной по итогам шага 1.
- *Шаг 3.* Рассчитываем сумму значений, полученных по итогам шага 2, по всем ПЕВ в данной страте.
- *Шаг 4.* Рассчитываем сумму оцениваемых суммарных значений ПЕВ из шага 1 по всем ПЕВ.
- *Шаг 5.* Результат шага 4 возводим в квадрат и делим на n_h, число ПЕВ в страте.
- *Шаг 6.* Результат шага 5 вычитаем из результата шага 3, и полученную разность умножаем на коэффициент $n_h/(n_h - 1)$. Это — оцениваемая вариантность характеристики на уровне страты.
- *Шаг 7.* Результат шага 6 суммируем по всем стратам для получения общей оцениваемой вариантности интересующей характеристики.
- *Шаг 8.* Рассчитываем квадратный корень из результата шага 7 для получения оцениваемой ошибки выборки для интересующей характеристики.

61. При расчете ошибки выборки для отношений, таких как оцениваемые доли, мы производим следующие действия:

- *Шаг 9.* Рассчитываем относительную вариантность числителя, $\hat{\hat{Y}}$, путем деления результата шага 7 на квадрат оценки числителя.
- *Шаг 10.* Повторяем шаг 9 для получения относительной вариантности знаменателя, $\hat{X}$.
- *Шаг 11.* Вычитаем результат шага 10 из результата шага 9.
- *Шаг 12.* Умножаем результат шага 11 на квадрат оцениваемого отношения $\hat{R}$. Это — расчетная вариантность $\hat{R}$.
- *Шаг 13.* Рассчитываем квадратный корень из результата шага 12 для получения оцениваемой ошибки выборки для $\hat{R}$.

Пример 4

Теперь рассмотрим гипотетический пример для иллюстрации метода кластера последнего этапа отбора при расчете вариантности. Предположим, что нас интересует оценка общих еженедельных расходов на продукты питания в домохозяйствах города А. Мы планируем обследование, применяя стратифицированную трехэтапную гнездовую выборку, включающую три страты с двумя ПЕВ, отбираемыми из каждой страты, и двумя домохозяйствами, отбираемыми из каждой включенной в выборку ПЕВ. Затем еженедельные расходы на продукты питания фиксируются по каждому домохозяйству, вошедшему в выборку обследования. В таблице 7.7 показаны данные такого обследования, включающие веса (W_{hij}) и еженедельные расходы на продукты питания, в долларах (Y_{hij}), для каждого вошедшего в выборку домохозяйства.

Таблица 7.7
Еженедельные расходы домохозяйств на продукты питания в разбивке по стратам

Страта	ПЕВ	Домашнее хозяйство	Вес (W_{hij})	Расходы в долларах (Y_{hij})	W_{hij}*Y_{hij}
1	1	1	1	30	30
		2	1	28	28
	2	1	3	12	36
		2	3	15	45
2	1	1	5	6	30
		2	5	7	35
	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
	2	1	4	13	52
		2	4	15	60
Всего			**42**		**474**

62. Согласно изложенному в главе 6, оценка общих еженедельных расходов на продукты питания для данного города определяется как

$$\hat{Y} = \sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}\, Y_{hij} = 474.$$

Кроме того, оценка средних еженедельных расходов на продукты питания определяется как

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}\, Y_{hij}}{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}} = \frac{474}{42} = 11 \text{ (округление до ближайшего целого доллара).}$$

63. Теперь мы осуществляем шаги по методу кластера последнего этапа отбора для расчета вариантности в столбцах таблицы 7.8, ниже. Числа в столбцах соответствуют указанным выше шагам.

Таблица 7.8
Осуществление шагов по методу кластера последнего этапа отбора для расчета вариантности

Страта	ПЕВ	Шаг 1	Шаг 2	Шаг 3	Шаг 4	Шаг 5	Шаг 6	Шаг 7
1	1	58	3 364	9 925	139	9 660,5	529	
	2	81	6 561	–	–		–	
2	1	65	4 225	8 849	133	8 844,5	9	
	2	68	4 624	–	–		–	
3	1	90	8 100	20 644	202	20 402	484	
	2	112	12 544	–	–		–	
Всего								1 022

64. Оценки вариантности на уровне страт составляют 529 для страты 1; 9 — для страты 2; и 484 — для страты 3. Общая оценка вариантности для расчетной величины еженедельного дохода домохозяйства (шаг 7 в нашей схеме) получается путем суммирования оценок на уровне страт, и эта сумма составляет 1 022.

7.7.3. Аппроксимации линеаризации

65. Большинство представляющих интерес оценок в обследованиях домашних хозяйств являются нелинейными. К некоторым из таких примеров можно отнести средний индекс массы тела для детей школьного возраста в той или иной стране; долю дохода, затрачиваемую на оплату жилья в данном городе; отношение вероятности наличия той или иной характеристики у одной подгруппы населения к вероятности наличия такой же характеристики у другой подгруппы населения и т. д. По методу линеаризации такие нелинейные показатели «линеаризуются» с помощью разложения в ряды Тейлора и последующего приближения вариантности оценки к вариантности первого порядка или линейной части этого разложения, используя точные методы, обсуждавшиеся в разделах выше.

66. Предположим, что мы хотим рассчитать вариантность оценки z параметра $\underline{Z}$, и предположим, что z — это нелинейная функция простых оценок $y_1, y_2, \ldots, y_m$ параметров $Y_1, Y_2, \ldots, Y_m$, а именно:

$$z = f(y_1, y_2, \ldots, y_m). \tag{7.24}$$

Исходя из того, что значение z близко к значению Z, разложение z в ряды Тейлора до первого порядка в z-$\underline{Z}$ выражается формулой

$$z = Z + \sum_{i=1}^{m} d_i (y_i - Y_i), \tag{7.25}$$

где d_i — это частичные производные от z в отношении y_i, иными словами,

$$d_i = \frac{\partial z}{\partial y_i},$$

что является функцией базисной оценки y_i. Это означает, что вариантность $\underline{z}$ может аппроксимироваться вариантностью линейной функции в приведенном выше уравнении (7.24), которое, как

мы знаем, рассчитывается с помощью точных методов, изложенных в предшествующих разделах, а именно:

$$v(z) = v(\sum d_i y_i) = \sum_{i=1}^{m} d_i^2 v(y_i) + \sum_{i \neq j} d_i d_j \operatorname{cov}(y_i, y_j). \quad (7.26)$$

67. Уравнение (7.26) содержит матрицу ковариантности (m × m) из m базисных оценок y_1, y_2, …, y_m, с m членами вариантности и $m(m–1)/2$ идентичными членами ковариантности, которые могут быть определены с помощью точных методов для линейной статистики, обсуждавшихся в предыдущих разделах.

Пример 5 (вариантность соотношения)

Для иллюстрации метода линеаризации рассмотрим расчет вариантности для соотношения

$$z = r = \frac{y}{x}. \quad (7.27)$$

Отметим, что для данного случая $\frac{\partial r}{\partial y} = \frac{1}{x}$ и $\frac{\partial r}{\partial x} = -\frac{y}{x^2} = -\frac{r}{x}$. Таким образом,

$$v(r) = \frac{1}{x^2}\left\{v(y) + r^2 v(x) - 2r \operatorname{cov}(y, x)\right\}, \quad (7.28)$$

что является известным выражением вариантности соотношения, которое можно найти в большинстве учебников по теории выборки.

68. Линеаризация широко используется на практике, поскольку она может применяться практически в любых планах выборки и для любых статистических показателей, которые поддаются линеаризации, т. е. которые могут быть выражены в виде линейной функции общеизвестных статистических показателей, таких как средние или суммарные значения, причем коэффициенты берутся из частичных производных, требуемых для разложения в ряды Тейлора. После линеаризации вариантность нелинейной оценки может быть аппроксимирована с помощью описанных выше точных методов (для ознакомления с техническими подробностями процесса линеаризации, включая соответствующие примеры см. работы Кокрана (1977 год) и Лора (1999 год)).

7.7.3.1. Преимущества

69. Поскольку метод расчета вариантности с помощью линеаризации используется уже в течение длительного времени, он имеет хорошо проработанную теоретическую базу и может применяться к более широкому кругу планов выборки по сравнению с применяемыми методами репликации (рассматриваемыми ниже). Если известны частичные производные, а квадратные члены и члены более высокого порядка при разложении в ряды Тейлора пренебрежимо малы, тогда линеаризация дает примерную оценку вариантности для практически всех представляющих интерес линейных алгоритмов оценки, таких как соотношения и коэффициенты регрессии.

7.7.3.2. Недостатки

70. Линеаризация дает достоверные результаты только в случае правильности приведенных выше допущений в отношении частичных производных и членов более высокого порядка. В противном случае в оценках могут появиться серьезные погрешности. Кроме того, обычно возникают затруднения при применении этого метода к сложным функциям с участием весов. Для каждого типа оценки должна разрабатываться отдельная формула, и это может потребовать создания специальных программ. Данный метод неприменим в отношении таких статистических показателей, как медианы и другие виды процентилей, не являющиеся гладкими функциями суммарных или средних значений совокупности.

71. Более того, в рамках метода линеаризации трудно применять поправки на неполучение ответов или неполный охват, которые зависят от плана выборки, представляющей интерес оценки и процедур взвешивания. Этот метод также требует включения в файл данных информации о плане выборки (страты, ПЕВ, веса).

7.7.4. Репликация

72. Репликационный подход включает целый класс методов, предусматривающих отбор из имеющихся данных повторных подвыборок или *реплик*, перерасчет взвешенной оценки обследования для каждой реплики и для полной выборки, а затем расчет вариантности как функции отклонений этих оценок репликаций от оценок полной выборки. Кратко этот подход можно описать в виде следующих шагов:

- *Шаг 1.* Удаляем различные подвыборки из полной выборки для формирования повторных выборок.
- *Шаг 2.* Получаем веса для реплик путем повторения процедуры расчета для каждой повторной выборки.
- *Шаг 3.* Получаем оценку из полной выборки и из каждого набора весов репликации.
- *Шаг 4.* Рассчитываем вариантность оценки в виде квадратов отклонения оценок репликаций от оценки полной выборки.

73. Предположим, например, что создаются k реплик из выборки, причем каждая реплика с оценками $\hat{\theta}_1$, $\hat{\theta}_2$, ... $\hat{\theta}_k$ параметра $\hat{\theta}$, и предположим, что оценкой, основанной на полной выборке, является $\hat{\theta}_0$. Тогда основанная на репликации оценка вариантности определяется как

$$\text{var}\ (\hat{\theta}) = \frac{1}{c}\sum_{r=1}^{k}(\hat{\theta}_r - \hat{\theta}_0)^2, \qquad (7.29)$$

где c — постоянная величина, которая зависит от метода расчета. Методы репликации отличаются в части значения постоянной величины и того способа, с помощью которого формируются реплики (краткий обзор наиболее часто используемых методов репликации см. в разделе 7.7.5).

7.7.4.1. Структура файла данных

74. Каким бы ни был метод репликации, структура файла данных остается неизменной, как показано в таблице 7.9, ниже.

Таблица 7.9
Структура файла данных для метода репликации

Запись	Данные	Вес полной выборки	Веса репликаций				
			1	2	3	….	k
1	Данные 1	w_1	w_{11}	0	w_{13}	…	w_{1k}
2	Данные 2	w_2	0	w_{22}	w_{23}	…	w_{2k}
3	Данные 3	w_3	w_{31}	w_{32}	0	…	w_{3k}
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
N	Данные *n*	w_n	w_{n1}	w_{n2}	w_{n3}	…	0

Примечание: Точки обозначают ряды данных

7.7.4.2. Преимущества

75. Основным преимуществом репликационного подхода по сравнению с линеаризационным подходом является то, что при репликации используется один и тот же базовый метод расчета вне зависимости от оцениваемого статистического показателя (поскольку оценка вариантности является функцией выборки, а не оценки), в то время как аппроксимация с помощью линеаризации должна разрабатываться аналитически для каждого статистического показателя, что потенциально является весьма трудоемким процессом в случае крупных обследований домашних хозяйств с большим числом представляющих интерес характеристик. Более того, методы репликации удобны в использовании и применимы практически ко всем статистическим показателям, линейным и нелинейным. При применении репликации оценки могут быть легко рассчитаны для подгрупп совокупности, при этом эффекты неполучения ответов и прочие поправки могут отражаться в весах репликаций.

7.7.4.3. Недостатки

76. Методы репликации требуют значительных вычислительных ресурсов, поскольку они требуют расчета набора весов репликаций, которые представляют собой аналитические веса, пересчитываемые для каждой из отобранных реплик, с тем чтобы каждая реплика, соответственно, представляла ту же совокупность, что и полная выборка. Кроме того, формирование реплик может осложняться ограничениями плана выборки (см. раздел 7.7.5, ниже), что иногда может приводить к завышенной оценке ошибок выборки.

77. Мы завершаем наше общее сравнение линеаризационного и репликационного подходов к расчету ошибок выборки замечанием, что эти два подхода не дают идентичных оценок ошибки выборки. Тем не менее эмпирические исследования (см. работу Киша и Франкеля, 1974 год) показали, что для крупных выборок и многих статистических показателей различия между результатами, полученными с помощью этих двух методов, пренебрежимо малы.

7.7.5. Некоторые методы репликации

78. Наиболее широко используемыми методами репликации являются следующие:

a) случайные группы;

b) сбалансированная многократная репликация (СМР);

c) репликация методом «складного ножа» (JK1, JK2 и JKn);

d) метод «расшнурованной выборки».

Ниже кратко рассматриваются поочередно каждый из этих методов.

7.7.5.1. Случайные группы

79. Метод случайных групп предусматривает разделение полной выборки на k групп таким образом, чтобы сохранить неизменным план выборки, иными словами, каждая группа представляет собой миниатюрную версию обследования, отражающую план выборки. Например, если полная выборка представляет собой ПСВ размера n, тогда случайные группы могут формироваться путем случайного распределения n наблюдений на k групп, каждая из которых будет иметь размер n/k. Если речь идет о гнездовой выборке, тогда ПЕВ в случайном порядке распределяются среди k групп таким образом, чтобы гарантировать, что в каждой ПЕВ будут сохранены все ее наблюдения, и, следовательно, каждая случайная группа будет по-прежнему представлять собой гнездовую выборку. Если речь идет о стратифицированной многоэтапной выборке, тогда случайные группы могут формироваться путем создания выборки ПЕВ из каждой страты. Отметим, что общее число случайных групп, которые предполагается сформировать в данном случае, не может превышать число вошедших в выборку ПЕВ из наименьшей страты.

80. Метод случайных групп может легко использоваться для оценки ошибок выборки как для линейных статистических показателей, таких как средние и суммарные значения и их гладкие функции, так и для нелинейных — таких как соотношения и процентили. Никакого специального программного обеспечения не требуется для оценивания ошибок выборки, которое является просто стандартным отклонением оценок, базирующихся на случайных группах, сформированных из полной выборки. Тем не менее создание случайных групп может быть затруднено в сложных планах выборки, поскольку каждая случайная группа должна сохранить структуру плана всего обследования в целом. Более того, число случайных групп может ограничиваться планом самого обследования. Например, при плане выборки с двумя ПЕВ в каждой страте могут быть сформированы только две случайные группы, при этом небольшое число случайных групп, как правило, приводит к неточной оценке ошибок выборки. Общее эмпирическое правило требует создания не менее 10 случайных групп для получения более стабильной оценки ошибок выборки.

7.7.5.2. Сбалансированная многократная репликация

81. Сбалансированная многократная репликация (СМР) предполагает наличие плана выборки с двумя ПЕВ на каждую страту. Формирование реплики предусматривает разделение каждой страты на две ПЕВ и отбор одной из этих двух ПЕВ в каждой страте в соответствии с предписанным планом для представления всей страты в целом. Этот метод можно адаптировать для других планов выборки путем группировки ПЕВ в псевдостраты по две ПЕВ в каждой.

7.7.5.3. Метод «складного ножа»

82. Как и СМР, метод «складного ножа» является обобщением метода случайных групп, который допускает наложение групп репликациий. Существуют три вида методов «складного ножа»: методы JK1, JK2 и JKn.

83. JK1 — это типовой метод «складного ножа» с удалением одной единицы для планов ПСВ. Тем не менее он может использоваться и для других планов, если вошедшие в выборку единицы объединены в случайные подгруппы, каждая из которых напоминает полную выборку.

84. JK2 аналогичен CMP в том плане, что он предполагает план с двумя ПЕВ на каждую страту. При наличии саморепрезентативных ПЕВ можно создавать пары вторичных единиц выборки (ВЕВ). Как и CMP, метод JK2 можно адаптировать для других планов путем объединения ПЕВ в псевдостраты по две ПЕВ в каждой. Затем одна ПЕВ в случайном порядке исключается поочередно из каждой страты в целях формирования реплик.

85. JKn — это типовой метод «складного ножа» с удалением одной единицы для стратифицированных планов выборки. Для создания реплик каждая ПЕВ исключатся поочередно из каждой страты. Оставшиеся в данной страте ПЕВ повторно взвешиваются для оценки суммарного значения страты. Число реплик равно числу ПЕВ (или псевдо-ПЕВ).

7.7.5.4. Метод «расшнурованной выборки»

86. Метод «расшнурованной выборки» начинается с отбора полной выборки, которая воспроизводит все важные характеристики совокупности в целом. Затем полная выборка рассматривается, как если бы она являлась всей совокупностью в целом, и из нее формируются подвыборки. Как и ранее, оценка ошибки выборки получается путем расчета стандартного отклонения оценок, базирующихся на повторных выборках, из оценок, основанных на полной выборке.

87. Метод «расшнурованной выборки» дает достоверные результаты с обычными планами выборки, а также с негладкими функциями, такими как процентили. При этом, однако, он требует большего объема вычислений, чем другие методы репликации.

88. В таблице 7.10, ниже, приводится значение постоянной величины c в формуле вариантности [уравнение (7.28)], которое соответствует различным методам репликации.

Таблица 7.10
Значения постоянного множителя в формуле вариантности для различных методов репликации

Метод репликации	Значение постоянной величины c в уравнении (7.28)
Случайная группа	$k(k–1)$
CMP	k
JK1	1
JK2	2
JKn	$k(k–1)$
Метод «расшнурованной выборки»	$k–1$

Пример 6 (метод «складного ножа»)

Теперь приведем простой численный пример применения метода «складного ножа» для расчета вариантности. Предположим, что у нас есть выборка размером 3. Мы можем создать 3 подвыборки, размером 2 каждая, путем единовременного исключения одной единицы из полной выборки. В таблице 7.11, ниже, представлены значения переменной (Y). Для трех подвыборок символ «X» обозначает, какие единицы выборки включены в подвыборку.

Вариантность выборки: $s^2 = \frac{(5-7)^2 + (7-7)^2 + (9-7)^2}{3-1} = 4.$

Оценка вариантности выборочного среднего значения (без учета fpc): $\frac{s^2}{n} = \frac{4}{3}.$

Таблица 7.11
Применение метода «складного ножа» по расчету вариантности к небольшой выборке и ее подвыборкам

		Подвыборка (g)		
Единица выборки	Y	1	2	3
1	5	X	X	
2	7	X		X
3	9		X	X
Общая величина выборки	**21**			
Выборочное среднее значение	7	6	7	8

Среднее из средних значений подвыборки: $\frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3}{3} = \frac{6+7+8}{3} = 7.$

Оценка вариантности среднего значения выборки по методу «складного ножа» определяется уравнением

$$V_J(\bar{y}) = \frac{n-1}{n}\sum_{g=1}^{3}(\bar{y}_g - \bar{y})^2 = \frac{3-1}{3}\left[(6-7)^2 + (7-7)^2 + (8-7)^2\right] = \frac{4}{3},$$

которое точно совпадает с уравнением для оценки вариантности выборочного среднего значения, расчет которого приведен выше.

Пример 7 (формирование реплик)

В таблице 7.12 использованы данные из примера 4 (раздел 7.7.2) для иллюстрации повторных выборок для различных методов репликации, а также приводится расчет значений вариантности методом «складного ножа».

Таблица 7.12
Полная выборка: расходы в разбивке по стратам

Страта	ПЕВ	Домашнее хозяйство	Вес W_{hij}	Расходы Y_{hij}	W_{hij}*Y_{hij}
1	1	1	1	30	30
1	1	2	1	28	28
1	2	1	3	12	36
1	2	2	3	15	45
2	1	1	5	6	30
2	1	2	5	7	35
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
3	2	1	4	13	52
3	2	2	4	15	60
Всего			**42**		**474**

Расчетное среднее значение, базирующееся на полной выборке:

$$\hat{\bar{Y}}_0 = \frac{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}\, Y_{hij}}{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}} = \frac{474}{42} = 11.$$

Таблица 7.13
Метод «складного ножа» (исключение ПЕВ 2 из страты 1)

Страта	ПЕВ	Домашнее хозяйство	Вес (W_{hij})	Расходы (Y_{hij})	W_{hij}*Y_{hij}
1	1	1	1	30	30
1	1	2	1	28	28
2	1	1	5	6	30
2	1	2	5	7	35
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
3	2	1	4	13	52
3	2	2	4	15	60
Всего			42		474

Расчетное среднее значение, базирующееся на указанной выше повторной выборке:

$$\hat{\bar{Y}}_1 = \frac{393}{36} = 11.$$

Мы можем продолжать этот процесс, поочередно исключая одну ПЕВ из каждой страты. Всего таким способом можно сформировать шесть повторных выборок. В таблице 7.14, ниже, показаны оценки средних еженедельных доходов домохозяйств, базирующиеся на каждой из шести повторных выборок.

Таблица 7.14
Оценки, основанные на репликах

Реплика j	Исключенная ПЕВ	Оценка $\hat{Y}_j$	$\hat{Y}_j - \hat{Y}_0$	$(\hat{Y}_j - \hat{Y}_0)^2$
1	ПЕВ 2, страта 1	11	0	0
2	ПЕВ 1, страта 1	10	-1	1
3	ПЕВ 2, страта 2	10	-1	1
4	ПЕВ 1, страта 2	12	1	1
5	ПЕВ 2, страта 3	12	1	1
6	ПЕВ 1, страта 3	13	2	4
Всего				8

По методу «складного ножа» оценка вариантности расчетного среднего значения определяется как:

$$\mathrm{var}_{JK}(\hat{\bar{Y}}) = \sum_{h=1}^{H}\left\{\frac{n_h - 1}{n_h}\sum_{j=1}^{n_h}(\hat{\bar{Y}}_j - \hat{\bar{Y}}_0)^2\right\} = \frac{1}{2}\times 8 = 4.$$

(Отметим, что для этого примера H=3, а n_h=2 для всех h.)

89. Мы завершаем этот раздел еще одним примером формирования повторных выборок с помощью метода сбалансированной многократной репликации. Показанные в таблице 7.15, ниже, результаты соответствуют порядку исключения ПЕВ, указанному в заголовке таблицы.

Таблица 7.15
Метод сбалансированной многократной репликации (Исключение ПЕВ 2 из страт 1 и 3; ПЕВ 1 — из страты 2)

Страта	ПЕВ	Домашнее хозяйство	Вес W_{hij}	Расходы Y_{hij}	W_{hij}*Y_{hij}
1	1	1	1	30	30
1	1	2	1	28	28
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
Всего			18		216

Расчетное среднее значение, базирующееся на указанной выше выборке по методу CMP:

$$\hat{\bar{Y}}_{1,BRR} = \frac{216}{18} = 12.$$

7.8. Сложности применения стандартных пакетов статистического программного обеспечения для анализа данных по результатам обследований домашних хозяйств

90. Надлежащий анализ данных по результатам обследований домашних хозяйств требует, чтобы оценки ошибок выборки рассчитывались методом, учитывающим трудности использования плана выборки, с помощью которого были получены эти данные. К таким трудностям относятся стратификация, формирование кластеров, разновероятностная выборка, неполучение ответов и различные корректировки весов выборки (подробную информацию по разработке и корректировке весов см. в главе 6). Стандартные статистические программные пакеты не позволяют учитывать эти усложняющие факторы, так как они, как правило, базируются на допущении, что элементы выборки были отобраны из совокупности путем простой случайной выборки. Как показано в главе 6, точечные оценки параметров совокупности находятся под воздействием весов выборки, связанных с каждым наблюдением. Эти веса зависят от вероятностей отбора и других характеристик плана выборки, таких как стратификация и формирование кластеров. Когда в них игнорируются веса выборки, эти стандартные пакеты дают искаженные точечные оценки. Проведение взвешенного анализа с помощью этих пакетов несколько снижает искажения в точечных оценках, но даже тогда ошибки выборки в точечных оценках часто оказываются весьма серьезно недооцененными, поскольку процедура оценки вариантности обычно не учитывает прочие особенности плана, такие как стратификация и формирование кластеров. Это означает, что выводы, сделанные на основе такого анализа, будут ошибочными. Например, различия между группами могут быть ошибочно признаны существенными, или могут быть ошибочно отклонены те или иные предположения. Неправильные выводы из анализа данных по результатам обследований

домашних хозяйств могут, например, оказать существенное влияние на распределение ресурсов и формирование политики на национальном и региональном уровнях.

91. Теперь мы используем пример из работы Брогана (2004 год) для иллюстрации того факта, что использование стандартных статистических программных пакетов может привести к искаженным точечным оценкам, чрезмерным стандартным ошибкам и доверительным интервалам и ошибочным критериям значимости. Этот пример базируется на комплекте данных, полученных в результате выборочного обследования охвата вакцинацией против столбняка, проведенного в Бурунди в 1989 году. Одной из целей этого обследования было сравнение серопозитивной реакции населения (определяемой, как уровень антител столбняка не менее 0,01 международных единиц на миллилитр (МЕ/мл) с историей проведения противостолбнячных вакцинаций. Дополнительную информацию о методологии обследования и его опубликованных результатах можно получить из работы Брогана (2004 год) и упомянутых в ней ссылках. В таблице 7.16, ниже, приведены оценки процентной доли женщин с серопозитивной реакцией и связанные с ними значения стандартной ошибки и доверительного интервала.

Таблица 7.16
Использование различных программных пакетов для расчета вариантности оценок обследования с указанием доли женщин, имеющих серопозитивную реакцию, среди недавно родивших женщин, Бурунди, 1988–1989 годы

Программный пакет	Доля женщин, имеющих серопозитивную реакцию	Стандартная ошибка	95-процентный доверительный интервал
SAS 8.2 MEANS без весов	74,9	2,1	(70,8; 79,0)
SAS 8.2 MEANS с весами	67,2	2,3	(62,7; 71,7)
SAS 8.2 SURVEYMEANS	67,2	4,3	(58,8; 75,6)
SUDAAN 8.0	67,2	4,3	(58,8; 75,6)
STATA 7.0	67,2	4,3	(58,8; 75,6)
EPI INFO 6.04d	67,2	4,3	(58,8; 75,6)
WESVAR 4.1	67,2	4,3	(58,8; 75,6)

92. Отметим, что точечные оценки являются одинаковыми для всех программ, где используются веса, однако имеется четкое различие между взвешенными и невзвешенными оценками. Более того, стандартные ошибки, полученные с помощью специального программного пакета, почти в два раза превышают ошибки, полученные с помощью стандартных программных пакетов, которые предполагают использование простой случайной выборки. Иными словами, стандартные программные пакеты существенно недооценивают вариантность оценок выборки. Это может иметь важные последствия для методики лечения. Например, если определенные меры воздействия планировались, исходя из доли женщин с серопозитивной реакцией в 65 процентов или ниже, тогда такие меры были бы реализованы на основе анализа, базирующегося на специальных программных пакетах, но они могли бы не состояться в случае анализа с помощью стандартных программных пакетов. В таблице 7.16 показано, что те программные пакеты, которые надлежащим образом определяют вариантность оценок обследования, дают примерно одинаковые результаты. В следующем разделе мы даем краткий обзор некоторых общедоступных статистических программных пакетов, которые используются для анализа данных по результатам обследований домашних хозяйств.

7.9. Программное обеспечение для оценки ошибок выборки

93. Рассмотренные выше методы оценки ошибок выборки уже в течение длительного времени используются в развитых странах и осуществляются в основном с помощью настраиваемых по индивидуальным требованиям компьютерных алгоритмов, которые разрабатываются государственными статистическими агентствами, научно-исследовательскими институтами и частными организациями, проводящими обследования. Последние достижения в компьютерных технологиях привели к разработке нескольких программных пакетов для применения этих методов. Многие из этих программных пакетов сегодня пригодны для персональных компьютеров. Такие программы используют лишь один из распространенных подходов к оценке вариантности, которые рассматриваются в разделе 7.7. Большинство этих программных пакетов предлагают наиболее широко используемые оценочные показатели, такие как средние значения, пропорции, соотношения и коэффициенты линейной регрессии. В некоторые пакеты включены также аппроксимации для широкого круга алгоритмов, таких как коэффициенты логистической регрессии.

94. В этом разделе мы приводим краткий обзор некоторых общедоступных компьютерных программ для оценки ошибок выборки по данным обследований домашних хозяйств. Это ни в коей мере не полный список всех имеющихся программ и программных пакетов. Мы ограничиваем рассмотрение несколькими статистическими программными пакетами, которые в настоящее время доступны для персональных компьютеров, в целях использования аналитиками общих данных обследования. Дается краткий обзор каждого программного пакета с указанием всех подходящих для него планов выборки и методов оценки вариантности. Перечисляются также достоинства и недостатки применения данных программ. Мы не пытаемся дать подробное описание технических и вычислительных процедур, содержащихся в этих пакетах. Такую информацию можно получить на веб-сайтах данных программных пакетов, а также из некоторых ссылок, указанных в конце этой главы.

95. К рассматриваемым здесь шести программным пакетам относятся: CENVAR, EPI INFO, PC CARP, STATA, SUDAAN и WESVAR. Все программные пакеты SUDAAN (Shah, Barnwell and Bieler, 1998), STATA (StataCorp, 1996), PC CARP (Fuller and others, 1989) и CENVAR используют метод линеаризации в целях оценки ошибок выборки для нелинейных статистических показателей. В программе WESVAR используются исключительно методы репликации. В последних версиях программы SUDAAN также применяются методы CMP и «складного ножа». Кроме того, для программ SAS и SPSS (которые здесь не рассматриваются) разработаны новые модули для анализа данных обследования. Программы, основанные на репликации, способны предложить многие базисные методы, за исключением метода «расшнурованной выборки». Мы даем лишь краткое сравнение общих особенностей этих программных пакетов. Детальное сопоставление потребует более широкого сравнения по выборочным обследованиям различных размеров и гораздо большего числа статистических показателей, что выходит за пределы ограниченных масштабов данного обзора.

96. Интернет-ссылки на некоторые статистические программные пакеты, рассматриваемые в данной главе, а также на многие другие программы, можно найти на следующем веб-сайте: www.fas.harvard.edu/~stats/survey-soft/survey-soft/

97. В работе Брогана (2004 год) на основе данных упомянутого выше обследования в Бурунди приводится подробное сравнение нескольких статистических программных пакетов, включая те, которые рассматриваются здесь.

98. Ниже дается краткий обзор программных пакетов. Заинтересованные читатели могут получить более подробную информацию из последних учебных пособий или указанных ниже веб-сайтов.

CENVAR

United Stats Bureau of the Census; contact International Programs Center
United States Bureau of the Census
Washington, D.C. 20233-8860
E-mail: IMPS@census.gov
Website: www.census.gov/ipc/www/imps

99. CENVAR является компонентом статистической системы программного обеспечения, разработанной Бюро переписей Соединенных Штатов для обработки, управления и анализа данных комплексных обследований, и известной как Интегрированная микрокомпьютерная система обработки данных (ИМСО). Эта система применима для большинства планов выборки, таких как простая случайная выборка, стратифицированная случайная выборка и многоэтапная гнездовая выборка. Для расчета вариантности в CENVAR используется аппроксимация линеаризации.

100. К получаемым с помощью CENVAR оценочным показателям относятся средние значения, пропорции и суммарные значения по всей выборке в целом, а также по указанным подклассам в формате таблицы. Наряду с ошибкой выборки программа предоставляет 95-процентные пределы доверительных интервалов, коэффициенты вариации, эффекты схемы и невзвешенные размеры выборки.

EPI INFO

United States Centers for Disease Control and Prevention
Epidemiology Program Office, Mailstop C08
Centers for Disease Control and Prevention
Atlanta, GA 30333
E-mail: EpiInfo@cdc1.cdc.gov
Website: www.cdc.gov/epiinfo/

101. EPI INFO — это статистическая система программного обеспечения, разработанная Центрами контроля и профилактики заболеваний Соединенных Штатов для обработки, управления и анализа эпидемиологических данных, включая данные комплексных обследований (модуль CSAMPLE). Соответствующая документация в онлайновом режиме доступна прямо в программе, при этом может быть распечатана каждая глава. Программа создана конкретно для стратифицированной многоэтапной гнездовой выборки по модели выборки кластера последнего этапа.

102. EPI INFO дает оценки ошибок выборки для средних значений и пропорций по всей выборке в целом, а также по подклассам, указанным в форме таблицы с двумя признаками. Распечатываемые результаты включают невзвешенные показатели частоты, взвешенные пропорции или средние значения, стандартные ошибки, 95-процентные пределы доверительных интервалов и эффекты схемы.

PC CARP

Iowa State University
Statistical Laboratory
219 Snedecor Hall
Ames, IA 50011
Website: http://cssm.iastate.edu/software/pccarp.html

103. PC CARP — статистический программный пакет, разработанный в Университете штата Айова в целях оценки стандартных ошибок для средних значений, пропорций, квантилей, разности коэффициентов и анализа таблиц сопряженности по двум признакам. Эта программа предназначена для обработки стратифицированной многоэтапной гнездовой выборки. Для оценки вариантности в PC CARP используется линеаризационный подход.

STATA

Stata Corporation
702 University Drive East
College Station, TX 77840
E-mail: stata@stata.com
Website: www.stata.com

104. STATA — это программный пакет в целях статистического анализа, предназначенный для оценки ошибок выборки для средних и суммарных значений, соотношений, пропорций, линейной регрессии, логистической регрессии и процедур пробит-анализа. Дополнительные возможности включают оценку линейных комбинаций параметров и критериев проверки гипотез, а также расчет квантилей, анализ таблиц сопряженности признаков, компенсацию отсутствующих данных и другие виды анализа. Для оценки вариантности в программе STATA используется линеаризационный подход.

SUDAAN

Research Triangle Institute
Statistical Software Center
Research Triangle Institute
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194
E-mail: SUDAAN@rti.org
Website: www.rti.org/patents/sudaan.html

105. SUDAAN — это статистический программный пакет для анализа коррелированных данных, включая данные комплексных обследований. Он применим к широкому спектру планов выборки, включая простую случайную выборку и многоэтапные стратифицированные планы. Программа имеет функции для оценки широкого круга статистических показателей и связанных с ними ошибок выборки, включая средние значения, пропорции, соотношения, квантили, комбинационные таблицы и отношения шансов; функции моделирования линейных, логистических регрессий и регрессий пропорциональных рисков; функции анализа таблиц сопряженности признаков. Для оценки вариантности эта программа использует линеаризационный подход.

WESVAR

Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850-3129
E-mail: WESVAR@westat.com
Website: www.westat.com/wesvar/

106. WESVAR — это статистическая система программного обеспечения, разработанная компанией «Westat, Inc.» для анализа данных комплексных обследований, включая анализ таблиц сопряженности признаков, регрессию и логистическую регрессию. Эта система применима к большинству планов выборки, но специально предназначена для стратифицированных многоэтапных гнездовых выборок, базирующихся на модели выборки кластера последнего этапа.

107. Для оценки вариантности в программе WESVAR используются методы репликации, включая метод «складного ножа», сбалансированный метод половинной выборки и модификацию Фея для сбалансированного метода половинной выборки. Это требует создания новой версии набора данных в особом формате WESVAR и детализации весов реплик.

7.10. Общее сравнение программных пакетов

108. Рассмотренные здесь программные пакеты имеют множество общих особенностей. Все программы требуют детализации весов, страт и единиц выборки для каждого элемента выборки. Не каждая из этих программ способна обрабатывать все возможные планы выборки без погрешностей. Например, первичные единицы выборки в большинстве стратифицированных многоэтапных планов выборки формируются с вероятностью, пропорциональной размеру, и без замещения. Только одна программа из вышеприведенного списка — SUDAAN — обладает функциями для обработки планов выборки исключительно этого типа. Тем не менее все перечисленные программы могут работать с таким планом с помощью модели отбора формирования выборки по методу кластера последнего этапа (см. раздел 7.7, выше). Более того, программа SUDAAN также имеет функции, цель которых — оценить вариантности для планов, в которых применяется отбор первичных единиц выборки без замещения. STATA является единственным программным пакетом с функциями проведения оценок для учета стратификации и многоэтапного отбора, предусмотренных таким планом.

109. Все программные пакеты дают оценку вариантности выборки и связанных с этим статистических показателей (эффекты схемы, внутриклассовая корреляция и т. д.) для средних и суммарных значений и пропорций для полной выборки, для подклассов полной выборки и для разности между подклассами. Большинство из них оценивают вариантность выборки для статистики регрессии и логистической регрессии. Все они выдают проверочные статистические данные, базируясь на получаемых оценках вариантности выборки.

110. Программы CENVAR, EPI INFO, PC CARP и WESVAR доступны либо бесплатно, либо по номинальной цене. Заинтересованным пользователям необходимо обратиться к адресам электронной почты и прочей контактной информации для получения сведений о том, как получить эти программы и сопутствующую документацию.

7.11. Заключительные замечания

111. В этой главе представлен краткий обзор процедур оценки ошибок выборки, базирующихся как на стандартных, так и на более сложных планах выборки, которые применяются в обследованиях домашних хозяйств. Расчет ошибок выборки является важнейшим аспектом анализа результатов обследований домашних хозяйств и составления отчетов по ним. В идеальном случае ошибки выборки должны рассчитываться для всех характеристик по комплекту таблиц обследования. На практике, однако, определяется набор ключевых вызывающих интерес характеристик, по которым ведется расчет ошибок выборки для каждой из областей обследования. Должны выбираться характеристики, важные для предмета обследования, но при этом они также должны включать в себя репрезентативный набор позиций, имеющих определенные статистические свойства, а именно те, по которым ожидается масштабное формирование кластеров (например, переменные величины, указывающие на этническую принадлежность или доступ к тем или иным услугам), а также те, по которым ожидается низкий уровень эффекта формирования кластеров (такие, как семейное положение). Кроме того, при этом выборе необходимо руководствоваться другими свойствами, такими как характеристики, объединяющие высокую или низкую долю населения или важные вызывающие интерес исследуемые области.

112. В этой главе также поддерживается использование специального программного обеспечения для оценки ошибок выборки по данным обследований. Мы привели примеры ситуаций, при которых допускаются серьезные погрешности в оценке ошибок выборки в случае применения стандартных статистических программных пакетов. Как правило, при применении стандартных ста-

тистических программных пакетов для анализа данных по результатам обследований домашних хозяйств получается заниженная величина истинной вариантности оценок обследования. Такие заниженные оценки стандартной ошибки могут привести к ложным выводам касательно результатов обследования за счет, например, ошибочного признания существенной разницы между средними значениями двух групп или некорректного отклонения того или иного предположения.

113. В этой главе также приводится каталог некоторых общедоступных статистических программных пакетов наряду с базовой контактной информацией и обзором возможностей их применения. Отсутствие знаний или опыта в оценке ошибок выборки является одним из препятствий для проведения сложного анализа данных в развивающихся странах. Многие аналитики не имеют представления о необходимости использования специализированного программного обеспечения или, если даже они осведомлены об этом, предпочитают не проходить подготовку по использованию какого-либо нового программного пакета.

114. Необходимо подчеркнуть, что данная глава представляет собой лишь введение в обширную и растущую сферу оценки вариантности данных по результатам сложных обследований. Читателям рекомендуется ознакомиться с некоторыми приведенными в конце главы ссылками для более детального и систематического изучения этой темы. Более широкий обзор этих и других программных пакетов, включая систему команд и выходную информацию по некоторым имеющимся программным пакетам, дается в работе Брогана (2004 год) и в приводимых в ней ссылках.

115. И наконец, следует признать, что в условиях быстрого технического прогресса многие программные пакеты за относительно короткий срок либо устаревают, либо подвергаются модификации по сравнению с параметрами, приведенными в этом обзоре. Действительно, существует возможность того, что некоторые спецификации устареют уже на момент публикации этого руководства. В связи с этим важно помнить, что на момент их использования следует получить самую точную информацию по программным пакетам из их соответствующих инструкций или веб-сайтов.

Справочная литература и дополнительные источники информации

An, A. and D. Watts (2001). New SAS procedures for analysis of sample survey data. *SUGI Paper,* No. 23, Cary, North Carolina: SAS Institute, Inc. Available from http://support.sas.com/rnd/app/papers/survey.pdf.

Binder D. A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review,* Vol. 51, pp. 279–92.

Brick J. M., and others J. (1996), *A User's Guide to WesVarPC*, Rockville, Maryland: Westat, Inc.

Донна Броган (2005). «Оценки ошибок, обусловленных выборкой, в данных обследования» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96, в продаже под № R.05.XVII.6.

Burt, V. L., and S. B. Cohen (1984). A comparison of alternative variance estimation strategies for complex survey data. *Proceedings of the American Statistical Association Survey Research Methods Section.*

Carlson, B. L., A. E. Johnson and S. B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data, *Journal of Official Statistics* 9, No. 4, pp. 795–814.

Dippo, C. S., R. E. Fay and D. H. Morganstein (1984). Computing variances from complex samples with replicate weights. *Proceedings of the American Statistical Association Survey Research Methods Section.*

Fuller, Wayne, and others (1989). PC CARP: USERS MANUAL, Ames, Iowa: Statistics Laboratory, Iowa State University. Available from http://cssm.lastate.edu/software.

Hansen, M. H., W. N. Hurwitz and W. G. Madow (1953). *Sample Survey Methods and Theory,* Vol. I, *Methods and Applications*. New York: Wiley, Sect. 10.16.

Hansen M. H., W. G. Madow and B. J. Tepping (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association vol.* 78, No. 384, pp. 776–793.

Kist, Leslie (1965). Survey Sampling. New York: John Wiley and sons.

_____(1995). Leslie Kish: Selected Papers, Steven Heeringa and Graham Kalton, eds. Hoboken, New Jersey: John Wiley & sons, Inc.

Kish, L., and M. R. Frankel (1974). Inference from complex samples. *Journal of the Royal Statistical Society: services* B, vol. 36, pp. 1–37.

Landis J. R., and others (1982). A statistical methodology for analyzing data from a complex survey: the First National Health and Nutrition Examination Survey. *Vital and Health Statistics, vol. 2, No. 92,* Washington, D.C.: Department of Health, Education and Welfare.

Lehtonen, R., and E. J. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley.

Lepkowski J. M., J. A. Bromberg, and J. R. Landis (1981), A program for the analysis of multivariate categorical data from Complex Sample Surveys. *Proceedings of the American Statistical Association Statistical Computing Section.*

Levy, Paul S., and Stanley Lemeshow (1999). Sampling of Populations: Methods and Applications. 3rd ed. New York: John Wiley & Sons.

Lohr, Sharon (1999). Sampling: Design and Analysis. Pacific Grove, California: Duxbury Press.

Potthoff, R. F., M. A. Woodbury, and K. G. Manton (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association, vol.* 87, pp. 383–396.

Rust K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys", *Journal of Official Statistics* 1(4), 381–397.

Rust, K. F., and J. N. K. Rao (1996). "Variance estimation for complex surveys using replication techniques", Statistical Methods in Medical Research, vol. 5, pp. 283–310.

Shah, Babhai V. (1998) Linearization methods of variance estimation. In Encyclopedia of Biostatistics, vol. 3, Peter Armitage and Theodore Colton, eds. New York: John Wiley and sons, pp. 2276–2279.

Shah B. V., B. G. Barnwell and G. S. Bieler, (1996). *SUDAAN User's Manual: Release 7.0*, , Research Triangle Park, North Carolina: Research Triangle Institute.

Tepping B. J. (1968), Variance estimation in complex surveys, *Proceedings of the American Statistical Association Social Statistics Section*, pp. 11–18.

United Nations (1993). Sampling errors in household surveys. UNFPA/UN/INT-92-P80-15E. New York: Statistical Division, Department for Economic and Social Information and Policy Analysis, and National Household Survey Capability Programme.

Wolter, K. M. (1985), *Introduction to Variance Estimation*. New York: Springer-Verlag.

Woodruff R. S. (1971), A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association, vol.* 66, No. 334, pp. 411–414.

Глава 8

Ошибки регистрации данных в обследованиях домашних хозяйств

8.1. Введение

1. Как ошибки выборки, так и *ошибки регистрации данных* необходимо контролировать и снижать до такого уровня, на котором их присутствие не препятствует применимости конечных результатов обследования. В предыдущих главах касательно планов выборки и методов расчета основное внимание было сосредоточено на ошибках выборки, и несколько меньший упор делался на прочих источниках вариантности в обследованиях, таких как неполучение ответов и неполный охват, которые составляют класс ошибок, в общем известных как *ошибки регистрации данных*. Ошибки регистрации данных причиняют особый вред, если они имеют неслучайный характер в силу той погрешности, которую они вносят в оценки обследования.

2. Любые данные обследований подвержены ошибкам из различных источников. Общее, фундаментальное различие связано с ошибками, возникающими в процессе измерения, и ошибками выборки, иными словами — ошибками в оценке показателей совокупности, возникающими из-за измерения выборки данной совокупности.

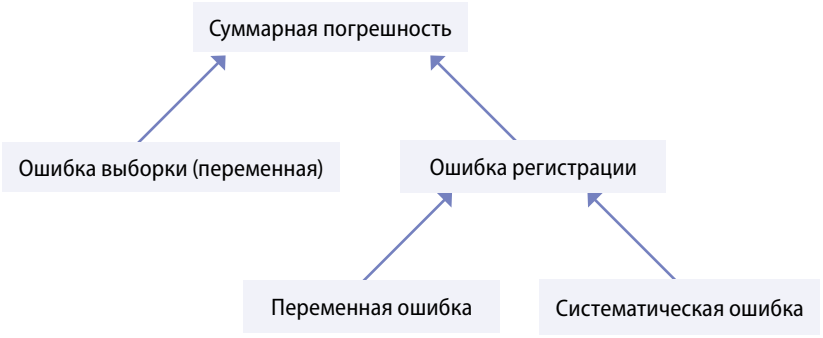
3. В предыдущих главах было сделано допущение, что каждая единица Y_i в совокупности ассоциируется со значением y_i, считающимся истинным значением единицы для характеристики y. Кроме того, было сделано допущение, что в любом случае, когда единица Y_i входит в состав выборки, полученное или наблюдаемое в ней значение y составляет y_i. Это справедливо в некоторых, но не во всех ситуациях. Например, в странах с надежной и всеобъемлющей системой регистрации этапов в жизни человека с помощью свидетельств о рождении, «истинные» значения могут быть легко получены, если y_i обозначает возраст. Однако в других ситуациях, предусматривающих качественную оценку состояния здоровья человека, истинные данные получить или даже определить бывает значительно сложнее. Например, больной человек может рассматривать себя здоровым в зависимости от обстоятельств.

4. Из практики проведения обследований ничем не оправдано предположение о том, что предоставленное в отчетности или наблюдаемое значение единицы Y_i всегда составляет y_i вне зависимости от того, кто сообщает эти данные или при каких обстоятельствах они получены. Фактический опыт проведения обследований дает множество примеров ошибок измерения или наблюдения, а также ошибок в результате неправильных ответов, неполучения ответов или по иным причинам в ходе любого обследования.

5. В дополнение к ошибкам в ответах, обследования подвержены ошибкам охвата, обработки данных и т. д. Качество выборочного алгоритма оценки какого-либо параметра совокупности является функцией суммарной погрешности обследования, состоящей как из ошибки выборки, так и из ошибки регистрации данных. Как уже указывалось ранее, ошибка выборки возникает исключительно в результате формирования вероятностной выборки, а не проведения поголовной

регистрации. Ошибка регистрации данных, с другой стороны, связана в основном с процедурами сбора и обработки данных. На рисунке 8.1 показана взаимосвязь между ошибками выборки и ошибками регистрации данных как элементами суммарной погрешности обследования.

Рисунок 8.1
Взаимосвязь между ошибками выборки и ошибками регистрации данных как составляющими элементами суммарной погрешности обследования



6. Таким образом, *ошибка регистрации данных* возникает в основном по причине ошибочных определений и понятий, неточных инструментариев выборки, вопросников неудовлетворительного качества, несовершенных методов сбора данных, составления таблиц и кодирования, а также неполного охвата единиц выборки. Эти ошибки не прогнозируются и их трудно контролировать. В отличие от *ошибки выборки* ошибка этого вида может возрастать по мере увеличения размера выборки. При отсутствии надлежащего контроля ошибка *регистрации данных* может нанести больше вреда, чем ошибка выборки, для крупномасштабных обследований домашних хозяйств.

8.2. Погрешность и переменная ошибка

7. Как показывает таблица 8.1, ошибки обследования можно классифицировать как переменные ошибки (V*E*) или погрешности. Переменные ошибки возникают прежде всего из ошибок выборки, хотя ошибки регистрации, появляющиеся в основном в результате обработки данных, таких как кодирование или присвоение ключей, также вносят свой вклад в переменную ошибку. Напротив, погрешности возникают в основном в результате ошибок регистрации данных в связи с такими факторами, как неправильные определения, ошибочные процедуры измерения, ошибочные ответы, неполучение ответов, неполный охват обследуемой совокупности и т. д. Некоторые виды погрешностей могут быть также обусловлены ошибкой выборки: они возникают в результате расчетов вариантности выборки с применением алгоритма оценки вариантности, который неправильно отражает план выборки, приводя к завышению или занижению ошибок выборки.

8. Погрешностями, как правило, обозначаются систематические ошибки, оказывающие влияние на любое обследование, проводимое в соответствии с конкретным планом выборочного

Таблица 8.1
Классификация ошибок обследования

Переменные ошибки	Ошибка выборки
	Ошибка регистрации
Погрешность	Ошибка выборки
	Ошибка регистрации

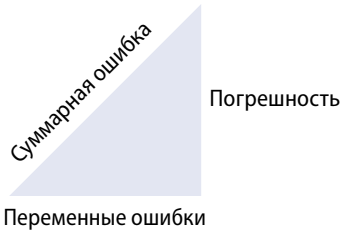
обследования с такой же постоянной ошибкой. Как отмечено в предыдущем пункте, ошибки выборки обычно являются причиной большинства переменных ошибок обследования, в то время как погрешности возникают в основном в результате *ошибок регистрации данных*. Таким образом, погрешности возникают из-за недостатков базисного плана и процедур обследования, в то время как переменная ошибка возникает по причине непоследовательности в применении планов и процедур обследования.

9. Статистическим термином для суммарной погрешности обследования выступает *среднеквадратическая ошибка* (*MSE*), которая равняется величине вариантности плюс квадрат погрешности (см. рисунок 8.2). Если, например, погрешность равняется нулю, то среднеквадратическая ошибка будет просто равна вариантности оценки. В обследованиях домашних хозяйств погрешность никогда не равняется нулю. Однако, как указывалось выше, измерение суммарной погрешности в обследованиях практически невозможно, частично в силу того, что ее расчет требует знания истинной величины совокупности, которая в большинстве случаев неизвестна. Источники погрешности так многочисленны и их природа столь сложна, что весьма редко предпринимаются попытки оценить их в сумме.

10. Треугольник на рисунке 8.2, ниже, обозначает суммарную ошибку и ее элементы. Высота треугольника обозначает погрешность, а основание — переменную ошибку. То, что гипотенуза обозначает величину суммарной ошибки, отражает концепцию того, что среднеквадратическая ошибка (то есть суммарная ошибка) равняется квадратному корню из суммы величин вариантности выборки и квадрата погрешности. Следовательно,

$$\textit{Среднеквадратическая ошибка} = \sqrt{VE^2 + Sesgo^2}\,. \qquad (8.1)$$

Рисунок 8.2
Суммарная погрешность обследования и ее составные элементы

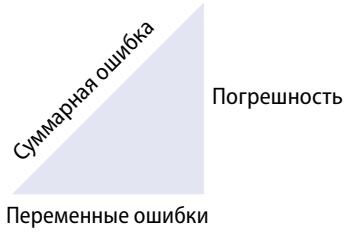


11. При уменьшении либо переменной ошибки, либо погрешности, соответственно уменьшается и суммарная ошибка. На рисунке 8.3 показана ситуация, когда значительно уменьшились и переменная ошибка, и погрешность. Вследствие этого существенно снизилась и суммарная ошибка, что следует из длины гипотенузы по сравнению с ее длиной на рисунке 8.2.

12. Одной из целей в рамках создания подходящего плана выборочного обследования наряду с эффективной стратегией проведения обследования является уменьшение как переменной ошибки, так и погрешности для получения сравнительно точных результатов выборки.

13. Как правило, высокая точность достигается за счет крупной и хорошо проработанной в других аспектах выборки, в то время как достоверности можно добиться только в случае сведения к минимуму (снижения) и переменной ошибки, и погрешности. Это предполагает, что точный план выборки может стать, тем не менее, весьма недостоверным, если он имеет высокую погрешность.

Рисунок 8.3
Уменьшенная суммарная ошибка обследования



В этом контексте важно помнить, что оценки стандартных ошибок, которые часто включаются в отчеты обследований домашних хозяйств, часто занижают суммарную ошибку обследования, поскольку эти оценки не учитывают влияние погрешностей.

14. На практике ошибки регистрации данных можно далее подразделить на переменные составляющие и систематические ошибки, см. работу Бимера и Либерга (2003 год). Систематические ошибки, как правило, представляют собой некомпенсирующиеся ошибки и поэтому имеют тенденцию к совпадению (чаще всего в одном и том же направлении), в то время как переменные ошибки являются компенсирующимися ошибками, которые имеют тенденцию к несовпадению (взаимноуничтожающие ошибки).

8.2.1. Переменная составляющая

15. Переменная составляющая ошибки возникает в результате случайных (произвольных) факторов, которые влияют на различные выборки и на повторения обследования. Применительно к процессу измерения мы можем представить, что весь спектр процедур — от выбора регистраторов и до сбора и обработки данных — можно повторять, используя одни и те же оговоренные процедуры при одних и тех же определенных условиях, а также независимо друг от друга в ситуациях, когда одно повторение не влияет на другое. На результаты повторений влияют случайные факторы, а также систематические факторы, возникающие в результате тех условий, при которых осуществляются такие повторения и которые одинаково влияют на результаты повторения.

16. Когда причиной появления переменных ошибок (*VE*) становятся только ошибки выборки, квадрат значений переменных ошибок равняется вариантности выборки. Отклонение среднего значения по результатам обследования от истинного значения для совокупности является погрешностью. И переменные ошибки, и погрешности могут возникать в результате либо процедур выборки, либо не связанных с выборкой операций. Переменная ошибка измеряет отклонение алгоритма оценки от его ожидаемого значения и включает как вариантность выборки, так и вариантность регистрации данных. Отличие ожидаемого значения алгоритма оценки от его истинного значения — это суммарная погрешность, включающая погрешности как обусловленные, так и не обусловленные выборкой.

17. Переменные ошибки могут оцениваться исходя из надлежащим образом разработанных критериев сравнения между повторениями (репликациями) тех или иных операций обследования при одинаковых условиях. Сокращение переменных ошибок зависит от выполнения более эффективного из двух условий — увеличения либо размера выборки, либо числа используемых регистраторов. С другой стороны, погрешность можно снизить только путем совершенствования процедур обследования, например введения мер контроля качества на различных этапах проведения обследования.

8.2.2. Систематическая ошибка (погрешность)

18. Систематическая ошибка возникает, например, в случаях, когда отмечается стабильная тенденция либо к занижению, либо к завышению оценок обследования. Например, в некоторых странах, где не существует свидетельств о рождении, среди мужчин наблюдается тенденция к указанию большего возраста, чем их фактический возраст. Такая практика со всей очевидностью приведет к систематической погрешности — завышению среднего возраста среди мужского населения.

8.2.3. Погрешность выборки

19. Погрешности выборки могут возникать в результате, как не соответствующего требованиям или ошибочного формирования заданной вероятностной выборки, так и ошибочных методов оценки. К первой группе относятся дефекты инструментариев, неправильные процедуры отбора, а также частичная или неполная регистрация отобранных единиц. В главах 3 и 4 данного руководства подробно рассматриваются разнообразные ситуации, при которых погрешности выборки могут возникать из-за неадекватного осуществления плана выборки — даже практически идеального.

8.2.4. Дальнейшее сравнение погрешности и переменной ошибки

20. Как правило, погрешности трудно поддаются измерению, и именно поэтому мы подчеркиваем необходимость их жесткого контроля. Их оценка может проводиться только путем сравнения результатов обследования с надежными внешними источниками данных. С другой стороны, переменную ошибку можно оценить с помощью сравнения подгрупп конкретной выборки или с помощью повторения обследования при тех же условиях. Уровень погрешностей можно снизить путем совершенствования процедур обследования.

21. По данным работы Вермы (1991 год), ошибки из некоторых источников — в том числе касающихся охвата, неполучения ответа и формирования выборки — возникают в основном в форме погрешностей. С другой стороны, ошибки кодирования и ввода данных могут появляться в основном в виде переменной ошибки.

22. Несмотря на то что и систематические, и переменные ошибки снижают общую точность и надежность, погрешности более негативно воздействуют на такие оценочные показатели, как средние значения, пропорции и суммарные значения для совокупности. Эти линейные показатели представляют собой сумму наблюдений в рамках выборки. Как уже отмечалось, переменные ошибки регистрации данных, как и ошибки выборки, могут быть снижены путем увеличения размера выборки. Для нелинейных оценочных показателей, таких как коэффициенты корреляции, стандартные ошибки и оценки регрессии, к серьезной погрешности могут привести как переменные, так и систематические ошибки (Biemer and Lyberg, 2003). Основной целью многих обследований домашних хозяйств является предоставление описательных характеристик совокупности, таких как средние и суммарные значения и пропорции для совокупности, вследствие этого основной упор в этих случаях необходимо делать на снижение систематической ошибки.

23. В заключение можно сказать, что погрешность возникает из-за недостатков базисного плана и процедур обследования. Ее более сложно измерить, чем переменную ошибку, и она может быть оценена только путем сравнения с более надежными источниками за пределами обычного обследования или с информацией, полученной при использовании более совершенных процедур.

8.3. Источники ошибок регистрации данных

24. Самые различные и весьма многочисленные причины ошибки регистрации данных присутствуют, начиная буквально с начального этапа, когда обследование еще только планируется и разрабатывается, и вплоть до завершающего этапа, на котором производится обработка и анализ данных.

25. Программа обследований домашних хозяйств может рассматриваться как свод строгих правил, оговаривающих различные операции. В этих правилах, например, дается описание подлежащей охвату целевой совокупности, оговариваются понятия и определения предмета обследования, которые будут использоваться в вопроснике, а также излагаются методы сбора данных и предполагаемые измерения. Если операция в рамках обследования проводятся в соответствии с установленными правилами, то существует теоретическая возможность получения *истинного значения* изучаемой характеристики. Однако ошибки регистрации данных делают это недостижимым идеалом.

26. Обычно ошибки регистрации данных могут возникать в результате наличия одного или нескольких из следующих факторов:

a) неадекватность и/или непоследовательность спецификации данных применительно к целям обследования;

b) дублирование или пропуск единиц в связи с неточным определением границ территориальных единиц выборки;

c) неполная или неточная идентификация реквизитов единиц выборки[1] или ошибочные методы ведения опроса;

d) ненадлежащие методы ведения опроса, наблюдения или измерения с применением допускающих двоякое толкование вопросников, определений или инструкций;

e) отсутствие подготовленных и опытных регистраторов на местах, включая отсутствие квалифицированного надзора на местах;

f) недостаточная проверка базовых данных для исправления явных ошибок;

g) ошибки, возникающие в ходе таких операций по обработке данных, как кодирование, присвоение ключей, проверка достоверности, составление таблиц и т. д.;

h) ошибки, возникающие в ходе презентации и публикации результатов в виде таблиц данных.

Тем не менее, приведенный список ни в коей мере не является исчерпывающим.

8.4. Элементы ошибки регистрации данных

27. В работе Бимера и Либерга (2003 год) указываются пять элементов ошибки регистрации данных, а именно ошибки спецификации, инструментария, неполучения ответов, измерения и обработки данных. Мы можем добавить ошибку оценивания в качестве еще одной ошибки, подлежащей рассмотрению. Краткий обзор этих ошибок приводится ниже.

1 Отметим, что хотя это происходит в рамках операции по формированию выборки, тем не менее — это один из видов погрешности регистрации данных.

8.4.1. Ошибка в спецификации

28. Ошибка в спецификации возникает, когда заложенное в вопросе понятие отличается от подразумеваемой и подлежащей измерению характеристики. Например, простой вопрос о том, сколько детей имеет тот или иной человек, может по-разному интерпретироваться в некоторых культурах. В домохозяйствах, где проживают расширенные семьи, биологические дети респондентов могут не отделяться от детей их братьев и сестер, проживающих в том же самом домохозяйстве. В обследованиях положения инвалидов вопрос общего характера о том, имеет ли человек инвалидность или нет, может интерпретироваться по-разному в зависимости от степени нетрудоспособности или восприятия респондентом самого понятия инвалидности. Люди с незначительной степенью нетрудоспособности могут не считать себя инвалидами. Если в вопросник не включить надлежащие проверочные и отсеивающие вопросы, ответы могут и не выявить общую численность лиц, имеющих инвалидность.

8.4.2. Ошибка охвата или инструментария

29. В большинстве территориальных обследований первичными единицами выборки являются географические единицы, такие как счетные участки переписи (СУ) (полномасштабное рассмотрение инструментариев выборки см. в главе 4). Довольно часто в ходе составления карт переписи демаркация границ СУ производится ненадлежащим образом. Вследствие этого в инструментарии выборки второго этапа определенные домохозяйства могут не учитываться или дублироваться. Такие недостатки могут вносить погрешности в оценки обследования по двум направлениям. Если единицы, которые должны присутствовать в инструментарии выборки там отсутствуют, то пропущенные единицы будут иметь нулевую вероятность отбора, что в результате приведет к недооценке. С другой стороны, в случае дублирования некоторых единиц результатом будут излишний охват и, следовательно, переоценка.

30. Таким образом, ошибки, связанные с инструментарием, могут в результате привести как к *избыточному охвату*, так и к *недостаточному охвату*. Последний из двух результатов является наиболее частым явлением в крупномасштабных обследованиях в большинстве африканских стран.

31. В многоэтапных обследованиях домашних хозяйств формирование выборки производится в несколько этапов, включая отбор территориальных единиц в один или более этапов; составление списков и отбор домохозяйств; составление списков и отбор лиц в вошедших в выборку домохозяйствах (см. главу 3). Ошибка охвата может возникнуть на любом из указанных этапов.

32. Важно еще раз подчеркнуть, что ни размеры, ни воздействие ошибок охвата оценить нелегко, поскольку такая оценка требует наличия внешней информации не только для самой выборки, но и, по определению, для используемого инструментария выборки.

33. Как отмечалось выше, *неполный охват* обозначает невключение в инструментарий выборки некоторых единиц определенной совокупности обследования (более подробное обсуждение вопроса об ошибке охвата, включая неполный охват, см. в главе 6). Поскольку такие единицы имеют нулевую вероятность отбора, они по существу исключаются из результатов обследования.

34. Важно отметить, что здесь не рассматривается преднамеренное и явное исключение сегментов более крупной совокупности из совокупности обследования. Такие преднамеренные исключения определяются целями обследования и возникающими в ходе него практическими трудностями. Например, из обследований отношения к браку могут исключаться лица, моложе минимального установленного законом возраста вступления в брак. Часто из-за сложностей обследования практи-

ческого характера исключаются лица, проживающие в специальных учреждениях. Районы страны с большим количеством заложенных противопехотных мин могут исключаться из обследований домашних хозяйств для обеспечения безопасности работников на местах. При расчете показателей неполного охвата члены преднамеренно и явно исключенной группы не должны учитываться ни в обследовании целевой совокупности, ни в показателях неполного охвата. В этой связи одним из четко заявленных необходимых условий обследования должно быть определение совокупности обследования (обсуждение вопроса целевой совокупности обследования см. в главе 3).

35. Термин *ошибка суммарного охвата* обозначает сумму абсолютных значений коэффициентов ошибок из-за *неполного охвата* и *избыточного охвата*. *Величина чистого недостающего охвата* обозначает превышение недостающего охвата над избыточным охватом. Таким образом, это — их алгебраическая сумма. Чистый охват измеряет полный охват, только если отсутствует избыточный охват. Большинство обследований домашних хозяйств в развивающихся странах страдают в основном от ошибок из-за неполного охвата. Большинство специалистов-практиков по изучению результатов обследований согласны с тем, что в большинстве социальных обследований неполный охват является значительно более распространенной проблемой, чем излишний охват. Поправки и взвешивание в случаях неполного охвата вносить гораздо сложнее, чем в случаях неполучения ответов, поскольку показатели охвата не могут быть получены из самой выборки, а лишь из внешних источников.

36. Причиной ошибок из-за неполного охвата может стать использование несовершенных инструментариев единиц выборки, как это подробно рассматривалось в главе 4. Если инструментарии не обновляются и для экономии времени или денег используются устаревшие инструментарии, это может привести к серьезным погрешностям. Например, если используемый для обследования домашних хозяйств старый список жилищных единиц не обновлялся с даты его первоначального формирования (этот срок может составлять 10 лет до начала текущего обследования), тогда вновь добавленные жилищные единицы в выбранном счетном участке не войдут в инструментарий жилищных единиц второго этапа. Аналогичным образом, покинутые жилищные единицы останутся в инструментарии как пропуски. В такой ситуации будет иметь место как пропуск единиц, относящихся к совокупности, так и включение единиц, не относящихся к совокупности.

37. Иногда не удается локализовать или посетить некоторые единицы выборки. Эта проблема также возникает при использовании неполных списков. Кроме того, погодные условия или плохая транспортная доступность иногда не позволяют добраться до некоторых жилищных единиц в течение оговоренного периода обследования.

38. Как упоминалось в главе 3, основополагающей целью выборочного обследования домашних хозяйств является получение объективных результатов, которые позволят сделать правильные выводы в отношении заданной целевой совокупности на основе наблюдений над единицами выборки. В связи с этим результаты обследования могут быть искажены, если уровень неполного охвата различается между географическими регионами и такими подгруппами населения, как мужчины/женщины, возрастные группы, этнические и социально-экономические классы.

39. Ошибки из-за неполного охвата отличаются от ошибок в связи с неполучением ответов. Последние, как обсуждалось ранее, возникают в результате неполучения наблюдений по некоторым единицам выборки вследствие отказов, невозможности найти адреса или обнаружить респондентов дома, утери вопросников и т. д. Уровень неполучения ответов измеряется исходя из результатов выборки путем сравнения сформированной выборки с фактически осуществленной выборкой. Как отмечалось выше, уровень неполного охвата, наоборот, можно оценить только какой-либо внешней проверкой по отношению к операциям обследования.

8.4.2.1. Ошибки осуществления выборки

40. Ошибка осуществления выборки означает потери и искажения в инструментарии выборки: например, ошибочное применение квот или процедур отбора. Другим примером является ненадлежащая подмена в ходе работ на местах отобранных домохозяйств другими — более доступными или готовыми к сотрудничеству.

8.4.2.2. Снижение ошибки охвата

41. Наиболее эффективным путем снижения ошибки охвата является улучшение качества инструментария путем исключения ошибочно присутствующих там единиц или дубликатов единиц. Наилучшим образом это достигается путем обеспечения надлежащего обновления устаревших инструментариев (подробное обсуждение этого вопроса см. в главе 4). Важно также обеспечить возможность легкого определения местонахождения территориальных единиц выборки и находящихся в них домохозяйств. Наилучшим образом это достигается с помощью проведения надлежащих работ по составлению карт при составлении первоначального инструментария, что обычно делается в ходе последней переписи населения и жилого фонда.

8.4.3. Неполучение ответов

42. Как неоднократно отмечалось в настоящем руководстве, неполучение ответов означает неспособность получить ответы от некоторых единиц выборки. В связи с этим выборочную совокупность можно представить как разделенную на две страты: одну, состоящую из всех единиц выборки, по которым получены ответы, и другую — из всех единиц выборки, по которым невозможно было получить ответы.

43. Чаще всего случаи неполучения ответа не распределяются равномерно по единицам выборки, а в значительной степени концентрируются в определенных подгруппах совокупности. В результате различий в долях неполучения ответов распределение реализованной выборки по подгруппам будет отличаться от первоначально сформированной выборки. Это отклонение, вероятнее всего, обусловит погрешность вследствие неполучения ответов в случае, если переменные показатели обследования также связаны с такими подгруппами.

44. Долю *неполучения ответов* можно точно оценить в том случае, если ведется подсчет всех удовлетворяющих критериям элементам, входящим в выборку. Доля *получения ответов* обследования определяется как отношение числа заполненных вопросников по единицам выборки к общему числу удовлетворяющих критериям[2] единиц выборки (см. также главу 6). Рекомендуемой практикой является отражение показателей неполучения ответов во всех открытых публикациях данных по результатам обследования, при этом такие данные являются обязательными для официальных обследований. Неполучение ответов может быть обусловлено отсутствием дома включенных в выборку лиц, их отказом от участия в обследовании или невозможностью такого участия в силу неспособности отвечать на вопросы. Неполучение ответов может также иметь место в силу утерянных списков/вопросников и невозможности проведения обследования в некоторых районах по причине погодных условий, факторов местности или нарушения требований безопасности. Все категории неполучения ответов относятся к респондентам, удовлетворяющим критериям, и из них должны исключаться не соответствующие критериям единицы, как указано в

2 В отношении некоторых отобранных в выборку единиц может выясниться, что они находятся за пределами охвата обследования и, следовательно, не подлежат включению в него, например незаселенные, определенные под снос или покинутые жилые помещения.

сноске 2. Например, в обследовании рождаемости изучаемым населением в отобранных *СУ* будут только женщины репродуктивных возрастных групп, что, естественно, исключает женщин, не входящих в эти возрастные группы, и всех мужчин.

45. Как отмечалось в главе 6, существуют два вида неполучения ответов: *неполучение ответов от единицы* и *неполучение ответов на отдельные вопросы*. *Неполучение ответов от единицы* означает, что информация вообще не получена от данной единицы выборки, в то время как *неполучение ответов на отдельные вопросы* относится к ситуации, когда получена некоторая, но не вся информация от данной единицы. Неполучение ответов на отдельные вопросы проявляется в виде пропусков в записях данных по ответившим единицам выборки и может быть обусловлено отказами, упущениями со стороны регистраторов или неспособностью дать ответ. На отказ потенциального респондента принять участие в обследовании может повлиять множество факторов, в том числе: отсутствие мотивации, недостаток времени, деликатный характер некоторых исследуемых вопросов и т. д. Гроувз и Коупер в своей работе (1995 год) предлагают ряд причин отказов, которые включают социальный контекст исследования, характеристики респондента, план обследования (включая нагрузку на респондента), характеристики регистратора и взаимодействие между регистратором и респондентом. Что касается неполучения ответов на отдельные вопросы, некоторые вопросы обследования могут смутить респондента, могут рассматриваться им как деликатные и/или неподходящие для заявленной цели. Регистратор может пропустить вопрос или не записать ответ на него. Кроме того, ответ может быть отвергнут в ходе редактирования данных.

46. Масштабы (полного) неполучения ответов от единицы, среди прочих факторов, служат показателем общего восприятия обследования, его сложности, организации и управления, а, следовательно, сложности, ясности и приемлемости конкретных пунктов вопросника и качества работы регистратора в обработке этих пунктов.

47. Неполучение ответов вносит погрешность в результаты обследования, которая может быть весьма серьезной в тех ситуациях, когда не ответившие единицы не являются «репрезентативными» по отношению к ответившим единицам, что обычно и имеет место. Неполучение ответов увеличивает как ошибку выборки за счет уменьшения размера выборки, так и ошибки регистрации данных.

48. Усилия по повышению доли ответивших единиц часто ведут к изменениям процедурного характера, предусматривающим выбор тех или иных операций в ходе обследования. Например, в целях повышения доли ответивших единиц в обследовании рождаемости в Замбии 1978 года в качестве регистраторов привлекались учителя-женщины, для того чтобы задавать вопросы по применению противозачаточных средств и т. д. Идея состояла в том, что в случае использования в качестве регистраторов молодых мужчин будет более высокая доля отказов в силу существующего для молодых мужчин запрета обращаться с вопросами в адрес женщин пожилого возраста, в особенности с вопросами по сексуальной тематике, включая использование противозачаточных средств.

49. На практике невозможно полностью исключить вероятность неполучения ответов, тем не менее, такую вероятность можно свести к минимуму с помощью методов убеждения, повторных посещений домохозяйств с отсутствующими дома респондентами и прочих методов. Более подробную информацию по решению проблемы неполучения ответов на отдельные вопросы в данных того или иного обследования см. в главах 6 и 9.

8.4.4. Ошибки измерения

50. Эти ошибки возникают, когда наблюдаются или могут быть измерены отклонения от фактических величин единиц выборки. В основном такие ошибки концентрируются в существенных информационных материалах обследования, таких как определение целей обследования, их трансформация в применимые на практике вопросы, а также получение, запись, кодирование и обработка ответов. Следовательно, такие ошибки влияют на точность измерения на уровне отдельных единиц.

51. Например, разработка на начальном этапе неправильных или вводящих в заблуждение определений и концепций касательно формирования инструментария и схемы вопросника приведет к неполному охвату и разнообразным интерпретациям различными регистраторами, в результате чего появятся неточности в собранных данных.

52. Еще одним источником этой ошибки являются некомпетентные инструкции работникам на местах. В некоторых обследованиях расплывчатые и нечеткие инструкции потребуют от регистраторов принятия самостоятельных решений при проведении работ на местах. Источником ошибки могут стать и сами регистраторы. Иногда может оказаться неверной информация, собранная по тому или иному пункту вопросника для всех единиц; причиной этого чаще всего выступает недостаточная подготовка работников на местах.

53. Распространенной проблемой измерения являются сообщения о возрасте опрашиваемого населения в Африке по причине добавления возраста и предпочтения определенных цифр при указании возраста. Причиной этого и других примеров ошибки измерения могут быть как респонденты, так и регистраторы, или и те и другие. В некоторых случаях может иметь место определенное взаимодействие между ними, обусловливающее рост таких ошибок. Кроме того, ошибки наблюдения могут быть вызваны дефектами измерительных устройств или недостатками методов измерения.

54. Респонденты могут вносить ошибки за счет следующих факторов:

- непонимание ими вопроса(ов) обследования;
- невнимательные и неправильные ответы по причине, например, отсутствия правильного понимания цели(ей) обследования; в частности, респонденты могут уделять недостаточно времени для обдумывания вопросов;
- желание «сотрудничать», отвечая на вопросы, даже если они не знают правильных ответов на них;
- преднамеренная склонность к неправильным ответам, например, в обследованиях по деликатным темам, таким как доходы или болезни, вызывающие общественное порицание;
- провалы памяти в случаях продолжительного отчетного периода, показательным примером чего может служить сбор информации по товарам кратковременного пользования в обследованиях расходов.

55. Кумулятивный эффект от различных ошибок, возникающих из разнообразных источников, может быть весьма существенным, поскольку ошибки из различных источников могут и не быть взаимонейтрализующимися. Конечным результатом таких ошибок может стать значительная погрешность.

8.4.5. Ошибки обработки данных

56. В ошибки обработки данных включаются, в частности:

- ошибки редактирования;
- ошибки кодирования;
- ошибки ввода данных;
- ошибки программирования.

57. Перечисленные выше ошибки возникают на этапе обработки данных. Например, при кодировании открытых вопросов, связанных с экономическими характеристиками, кодировщики могут отклоняться от предписанных в соответствующих пособиях процедур кодирования и вследствие этого присваивать неправильные коды видам деятельности.

8.4.6. Ошибки расчета

58. Ошибки расчета возникают, прежде всего, по причине применения неправильных формул при расчете весов обследования. Ошибки могут также появляться в результате ошибочного расчета весов даже при применении правильной формулы. Оценки вариантности выборки (ошибки выборки) возникают, когда используемый алгоритм оценки вариантности не соответствует фактическому плану выборки, создавая таким образом ошибки в доверительных интервалах, связанных с точечными оценками обследования. В каждом из этих случаев результаты содержат погрешности.

8.5. Оценка ошибок регистрации данных

59. Как неоднократно указывалось в настоящей главе, источники ошибки регистрации данных многочисленны и разнообразны. Вследствие этого практически невозможно оценить суммарное значение ошибок регистрации данных, возникающих в ходе обследования. Тем не менее существует возможность изучения и оценки некоторых элементов ошибки регистрации данных, и этот вопрос обсуждается ниже.

8.5.1. Проверки соответствия данных

60. При разработке инструментов обследования (вопросников) особое внимание следует уделять включению в них определенных вспомогательных информационных пунктов, которые послужат в качестве проверочных индикаторов качества подлежащих сбору данных. Если сведения по таким дополнительным информационным пунктам получить легко, они могут собираться по всем охваченным обследованием единицам; в противном случае они могут собираться только по подвыборке единиц.

61. Например, в ходе обследования, проводимого после регистрации в рамках переписи, в котором использовался метод регистрации юридического населения, бывает целесообразно также собирать информацию по методу регистрации фактического населения для получения возможности рассчитать число временно присутствующих и временно отсутствующих лиц. Сравнение этих двух показателей позволит иметь определенное представление о качестве данных. Аналогичным образом, для оценки качества данных обследования может быть полезно включение в вопросник пунктов, дающих в итоге определенные сравнительно стабильные соотношения, такие как соотношения полов.

62. На этапе обработки данных обследования должны также использоваться проверки соответствия данных. Могут вводиться перекрестные проверки для обеспечения, например, того, чтобы лица, закодированные как главы домохозяйств, были не моложе определенного заранее оговоренного возраста, или чтобы рожавшие женщины не были моложе, скажем, 13 лет.

8.5.2. Проверка/установление правильности выборки

63. Одним из путей оценки и контроля некоторых видов ошибок регистрации данных в обследовании является независимое дублирование работы на различных ее этапах в целях содействия обнаружению и исправлению ошибок. По причинам практического характера такое проверочное дублирование может осуществляться только по какой-то выборке участков работы с использованием меньшей по численности группы хорошо подготовленного и опытного персонала. Если план выборки составлен правильно и если проверочная операция проведена эффективно, появляется возможность не только выявить ошибки, но и получить представление об их размерах. Если бы существовала возможность полной проверки работ в рамках обследования, значительно повысилось бы качество конечных результатов.

64. Что касается проверки самой выборки, то корректировочные работы могут проводиться только в отношении уже проверенной выборки. Негативный эффект от такого ограничения может быть несколько снижен путем разделения результатов на различных этапах обследования — т. е. заполненных списков, закодированных списков, таблиц расчетов и т. д. — на отдельные группы и проверки выборок в каждой из таких групп. В случаях, когда величина ошибок в какой-либо группе превышает оговоренный уровень, вся эта группа может быть проверена на предмет выявления и исправления ошибок, улучшая таким образом качество конечных результатов.

8.5.3. Проверки, проводимые после обследования, или проверки путем повторного опроса

65. Один из важнейших видов проверки выборки, который может использоваться для оценки ошибок в ответах, состоит в формировании подвыборки (или выборки в случае переписи) и проведения повторной регистрации силами более подготовленных и опытных сотрудников, чем персонал, который был занят на этапе основной регистрации. Для обеспечения эффективности такого подхода необходимо обеспечить, чтобы

- повторная регистрация проводилась немедленно после основного обследования для сведения к минимуму ошибки припоминания;
- были предприняты шаги к минимизации *эффекта модификации выборки по заданным условиям*, который может оказать основное обследование на работу по последующей проверке.

66. Как правило, проверочное обследование предназначено для облегчения оценки как *ошибок охвата*, так и *ошибок содержания*. В этих целях желательно в первую очередь провести повторную регистрацию всех единиц выборки на более высоких этапах, т. е. *СУ* и деревень, для выявления ошибок охвата с последующим повторным обследованием только конечных единиц выборки, обеспечивая таким образом надлежащую репрезентативность для различных частей совокупности, имеющих особую значимость с точки зрения ошибок регистрации данных.

67. Важнейшим преимуществом проверочного обследования является то, что оно облегчает единую проверку, состоящую, во-первых, из сопоставления данных, полученных в ходе этих двух

регистраций для единиц, вошедших в проверочную выборку, и, во-вторых, анализа выявленных индивидуальных различий. При обнаружении расхождений необходимо приложить усилия к определению их причины и более детально проанализировать характер и виды ошибок регистрации данных.

68. Если такая единая проверка невозможна в силу временных и финансовых ограничений, можно использовать альтернативную, но менее эффективную процедуру, называемую агрегированной проверкой. Этот метод предусматривает сравнение оценок параметров, полученных из данных проверочного обследования, с оценками для основного обследования. Агрегированная проверка дает лишь общее представление о чистой ошибке, являющейся результатом ошибок в сторону и увеличения, и уменьшения оценок. Единая проверка, напротив, дает информацию как по чистой, так и по общей ошибке.

69. В проверке после проведения обследования необходимо применять такие же концепции и определения, как и использованные в первоначальном обследовании.

8.5.4. Методы контроля качества

70. Существует весьма обширная сфера применения статистических методов контроля качества к работе по проведению обследования в силу крупных масштабов и периодического характера процедур, используемых в такой работе. В крупномасштабных обследованиях для оценки качества данных и повышения точности конечных результатов могут использоваться графики контроля и методы выборочного приемочного контроля. В качестве примера, 100 процентов работы любого сотрудника по вводу данных могут проверяться на первоначальном этапе его работы, и если показатель ошибок такого сотрудника находится ниже какого-то оговоренного уровня, то после этого может проверяться точность лишь выборочной части его работы.

8.5.5. Изучение ошибок припоминания

71. Как указывалось ранее в этой главе, ошибки в ответах возникают в связи с целым рядом фактором, таких как:

- отношение респондента к обследованию;
- метод проведения опроса;
- навыки регистратора;
- ошибка припоминания.

72. В этом списке *ошибка припоминания* требует особого внимания, поскольку она создает специфические проблемы, которые часто неподконтрольны респонденту. Она связана с продолжительностью отчетного периода и сроком, прошедшим между отчетным периодом и датой обследования. Последняя проблема может решаться путем выбора в качестве отчетного периода какого-то приемлемого временного интервала, предшествующего дате обследования или периода времени, по возможности близкого к этому интервалу.

73. Одним из способов изучения ошибки припоминания является сбор и анализ данных, связанных с более чем одним отчетным периодом, по выборке или подвыборке охваченных обследованием единиц. Основной проблемой такого подхода является проявление в той или иной степени *эффекта модификации выборки по заданным условиям*, возможно по причине влияния данных по одному отчетному периоду на данные, предоставляемые по другому периоду. Для предотвраще-

ния этого эффекта данные по разным рассматриваемым отчетным периодам могут собираться по различным единицам выборки. Следует отметить, что для такого сравнения необходимы большие выборки.

74. Другим подходом является сбор некоторой дополнительной информации, которая позволит получить оценки по различным отчетным периодам. Например, в демографическом обследовании можно собирать данные не только по возрасту респондента, но также и по дате, месяцу и году рождения. Те или иные расхождения покажут наличие любой ошибки припоминания применительно к сообщенному возрасту.

8.5.6. Взаимопроникающие подвыборки

75. Метод взаимопроникающих подвыборок предусматривает формирование общей выборки из двух или более подвыборок, которые должны отбираться идентичным образом, с тем чтобы каждая из них была способна дать достоверную оценку того или иного параметра совокупности. Этот метод помогает дать оценку качества информации, поскольку взаимопроникающие подвыборки могут использоваться для получения информации об ошибках регистрации данных, такой как расхождения, возникающие из-за различных погрешностей работы регистраторов, разные методы получения информации в ходе опроса и т. д.

76. После обследования подвыборок различными группами регистраторов и обработки данных по ним разными группами сотрудников на этапе составления таблиц сравнение оценок по различным подвыборкам обеспечит общую проверку качества результатов обследования. Например, если при сравнении оценок на базе четырех подвыборок, обследованных и обработанных различными группами проводящих обследование сотрудников, окажется, что три оценки имеют близкие значения друг с другом, а четвертая оценка сильно отличается от них, и что различие между ними превышает значение, которое может быть обоснованно приписано ошибке выборки, то тогда следует поставить под вопрос качество работы по такой отличающейся подвыборке.

8.6. Заключительные замечания

77. Ошибкам регистрации данных в выборочных обследованиях домашних хозяйств необходимо уделять надлежащее внимание, поскольку при отсутствии контроля они могут стать причиной весьма больших погрешностей в результатах обследований. В большинстве обследований очень мало внимания уделяется контролю таких ошибок с риском получения результатов, которые могут оказаться ненадежными. Наилучшим способом контроля ошибок регистрации данных является соблюдение надлежащих процедур во всех мероприятиях в ходе обследования, начиная с планирования и формирования выборки и вплоть до анализа результатов. В частности, стандартной практикой должна стать тщательная и интенсивная подготовка персонала для работ на местах, а вопросы обследования, особенно те, которые не прошли проверку в предыдущих обследованиях, должны проходить полномасштабное предварительное тестирование.

Справочная литература и дополнительные источники информации

Biemer, P., and L. Lyberg (2003). *Introduction to Survey Quality.* Wiley Series in Survey Methodology. Hoboken, New Jersey: Wiley.

Biemer, P. and others, eds. (1991). *Measurement Errors in Surveys.* Wiley Series in Probability and Mathematical Statistics, New York: Wiley.

Cochran, W. (1963). *Sampling Techniques,* New York: Wiley.

Groves, R., and M. Couper (1995). Theoretical motivation for post-survey non-response adjustment in household surveys, *Journal of Official Statistics,* Vol. 11, No. 1, pp. 93–106.

Groves, R., and other eds. (2000). *Survey Non-response*, Wiley-Interscience Publication. New York: John Wiley & Sons, Inc.

Hansen M., W. Hurwitz and M. Bershad, (2003). Measurement Errors in Censuses and Surveys. *Landmark Papers in Survey Statistics*, IASS Jubilee Commemorative Volume.

Kalton, G., and S. Heeringa. (2003). *Leslie Kish: Selected Papers.* Wiley Series in Survey Methodology, Hoboken, New Jersey: Wiley.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Murthy, M. (1967). *Sampling Theory and Methods.* Calcutta, India: Statistical Publishing Society.

Onsembe, Jason (2003). *Improving data quality in the 2000 round of population and housing censuses*, Addis Ababa, Ethiopia: UNFPA Country Technical Services Team.

Raj, D. (1972). *The Design of Sample Surveys*. New York: McGraw-Hill Book Company.

P. Chandhok (1998). *Sample Survey Theory*. London: Narosa Publishing House.

Shyam, U. (2004). *Turkmenistan Living Standards Survey 2003, Technical Report,* National Institute of State Statistics and Information, Ashgabad, Turkmenistan.

Som, R. (1996). *Practical Sampling Techniques*. New York: Marcel Dekker Inc.

Sukhatme, P. and others (1984). *Sampling Theory of Surveys with Applications*. Ames Iowa and New Delhi: Iowa State University Press and Indian Society of Agricultural Statistics.

United Nations (1982). *National Household Survey Capability Programme: nonsampling errors in household surveys: Sources, assessment and control*. DP/UN/INT-81-041/2. New York: United Nations, Statistics Division.

Verma, V. (1991). *Sampling Methods: Training Handbook.* Tokyo Statistical Institute for Asia and the Pacific.

Whitfold, D. and J. Banda (2001). *Post Enumeration Surveys: Are they Worth it?* United Nations Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid–Decade Assessment and Future Prospects, New York, 7–10 August.

Глава 9

Обработка данных по результатам обследований домашних хозяйств

9.1. Введение

1. В данной главе рассматривается обработка данных применительно к национальным обследованиям домашних хозяйств. Она начинается с описания цикла типового обследования домашних хозяйств, а затем рассматривается подготовка к обработке данных как неотъемлемая часть процесса планирования обследования.

2. За последние два десятилетия или около этого быстро развивались информационные технологии. Этот прогресс в свою очередь оказал серьезное воздействие на методы разработки и реализации статистических систем обработки данных по результатам обследований.

3. Основным событием в аппаратурном обеспечении стал переход от больших вычислительных систем к персональным компьютерным платформам. Персональные компьютеры становятся все более мощными с точки зрения как быстродействия, так и объема памяти. Сегодня персональные компьютеры могут осуществлять различные виды обработки данных в связи со статистическими операциями, начиная с мелкомасштабных обследований и заканчивая крупномасштабными статистическими операциями, такими как переписи населения и жилого фонда с весьма крупными выборками.

4. Параллельно с разработками аппаратурных средств произошло значительное улучшение качества программного обеспечения для обработки, анализа и рассылки статистических данных. В результате этого появилась возможность передачи некоторых задач по обработке данных от программистов к профильным специалистам.

5. Со временем появился целый ряд программных пакетов для обработки данных по результатам статистических обследований. Сильные стороны каждого из этих программных продуктов соотносятся с требованиями каждого из различных этапов обработки данных. Добавление к этой главе может служить в качестве справочного пособия по выбору программного обеспечения для различных этапов обработки данных обследования; в нем также дается краткое описание каждого из программных продуктов, упоминаемых в настоящей главе.

9.2. Цикл обследования домашних хозяйств

6. На рисунке 9.1 показан цикл типового обследования домашних хозяйств. В принципе все обследования проходят через один и тот же цикл со следующими типовыми этапами:

- *Планирование обследования*: Разработчики обследования принимают решения касательно основных целей и сфер применения результатов обследования, его основных итоговых показателей и основных вводных данных, процедур получения вводных данных (разработка и подготовка вопросника и связанных с ним инструментов обследова-

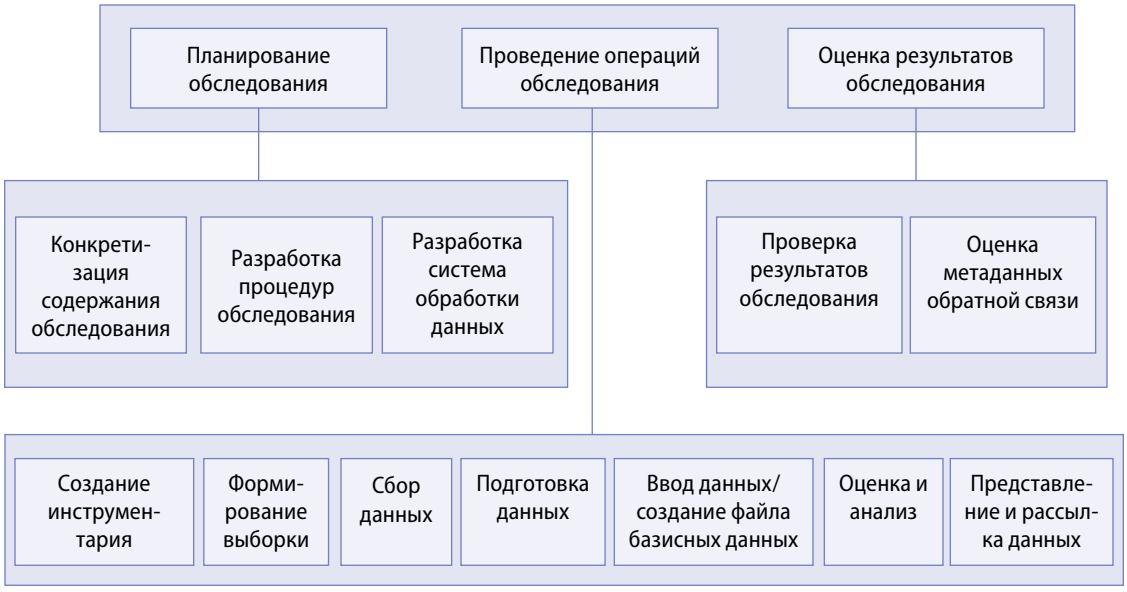
ния) и преобразования их в итоговые показатели, а также касательно проектирования системы обработки и документирования данных.

- *Проведение операций обследования*: Эти операции включают создание инструментария выборки, разработку плана и формирование выборки, сбор данных (измерение), подготовку данных (ввод, кодирование, редактирование и вменение данных), создание файла наблюдений (с базисными необработанными данными), осуществление расчетов, включая вычисление весов, создание производных переменных, анализ данных, а также представление и рассылку результатов.
- *Оценка результатов обследования*: Этот этап включает проверку и оценку того, получены ли оговоренные конечные продукты, были ли надлежащим образом опубликованы и прорекламированы итоговые результаты, были ли задокументированы и помещены в архив метаданные и т. д.

7. До того как приступить к разработке и реализации системы обработки данных для какого-либо конкретного обследования, важно наглядно представить себе общую систему для этого конкретного обследования. Надлежащая последовательность операций и процедур также играет весьма важную роль для успешного проведения любого обследования домашних хозяйств. Цели обследования должны определять в желаемой последовательности формат представления итоговых результатов (например, план составления таблиц и баз данных). Это, в свою очередь, будет определять последующие мероприятия по планированию обследования, сбору данных, подготовке и обработке данных и, в конечном счете, по анализу и рассылке результатов.

8. Обработка данных может рассматриваться как процесс, с помощью которого ответы на вопросы обследования, полученные в ходе сбора данных, переводятся в форму, подходящую для составления таблиц и анализа данных. Это предполагает смесь операций, производимых как автоматически, так и вручную. Этот процесс также может потребовать больших затрат времени и ресурсов и повлиять на качество и стоимость итоговых показателей.

Рисунок 9.1
Цикл обследования домашних хозяйств



Источник: Sundgren (1991).

9.3. Планирование обследований и система обработки полученных данных

9.3.1. Цели и содержание обследований

9. Как рассматривалось в главе 2, в качестве первого шага при планировании любого обследования необходимо сформулировать и документально оформить его основные цели. Обследования домашних хозяйств предоставляют информацию о характеристиках домохозяйств населения. Они проводятся для получения ответов на вопросы, которые могут возникать у тех или иных заинтересованных сторон касательно определенной части обследуемого населения. Поскольку цели конкретного обследования отражаются в тех случаях, когда делаются попытки получить ответы на такие вопросы, соответствующие данные должны предоставляться вопросником этого обследования.

10. Обычно те вопросы, на которые с помощью данных обследования домашних хозяйств хотят получить ответы заинтересованные стороны, могут разделяться на целый ряд категорий (см. работу Глевве, 2003 год).

11. С помощью одного набора вопросов следует установить основополагающие характеристики обследуемого населения (*доля населения, живущая в бедности, уровень безработицы и т. д.*).

12. Другой набор вопросов предназначен для оценки влияния мер вмешательства или общего развития событий на характеристики домохозяйств (например, *доля домохозяйств, участвующих в той или иной конкретной программе, сравнение их характеристик с характеристиками домохозяйств, не участвующих в данной программе, улучшение или ухудшение жилищных условий домохозяйств с течением времени и т. д.*).

13. И наконец, существует категория вопросов об определяющих факторах, или взаимосвязи, условий жизни домохозяйств и их характеристик (иными словами, *вопросы о том, что происходит и почему это происходит*).

9.3.2. Процедуры и инструменты обследований

9.3.2.1. Планы составления таблиц и ожидаемые результаты

14. Полезным методом для оказания помощи разработчику обследования в точном определении потребностей пользователей в информации является разработка планов составления таблиц и макетов таблиц. Макеты таблиц представляют собой проекты таблиц, содержащие все их элементы, кроме фактических данных. Как минимум, схема табулирования должна включать названия таблиц и заголовки столбцов, а также обозначать основные переменные, подлежащие включению в таблицы, фоновые переменные, используемые для классификации, группы совокупности (объекты, элементы или единицы обследования), к которым относятся различные таблицы (см. главу 2). Желательно также максимально подробно указать категории классификации, хотя они могут быть скорректированы позднее, когда станет лучше известно распределение выборки по категориям ответов.

15. Важность плана составления таблиц может рассматриваться с нескольких точек зрения. Составление макетов таблиц укажет, позволят ли собираемые данные составить пригодные для использования таблицы. Они не только укажут на то, каких данных не хватает, но и покажут излишние данные. Более того, дополнительное время, затраченное на составление макетов таблиц, обычно с лихвой компенсируется на этапах табулирования данных, сокращая время, затрачиваемое на разработку и составление фактических таблиц.

16. Существует также тесная взаимосвязь между планом составления таблиц и планом выборки, используемым для обследования. Например, географическая разбивка в таблицах возможна только в том случае, если выборка допускает такую разбивку.

17. Публикация Организации Объединенных Наций (1982 год) дает более подробное описание того, что включает в себя план составления таблиц и его различные преимущества.

18. Цель предыдущих пунктов — подчеркнуть ту важную роль, которую может сыграть план составления таблиц для эффективного планирования конкретного обследования и для соответствующей системы обработки данных. Важно подчеркнуть, однако, что план составления таблиц представляет собой лишь набросок некоторых итоговых показателей, которых можно ожидать от соответствующего обследования. Обследования домашних хозяйств имеют возможности предоставлять огромный массив информации. Обработанный набор микроданных обследования может рассматриваться как основной и базовый результат. Такой набор микроданных зачастую требует соответствующей «упаковки» и предоставления заинтересованным сторонам в удобном для них формате и по надлежащим каналам распространения.

9.3.2.2. Разработка и печатание формуляров

19. После того как определены цели обследования и план составления таблиц, можно приступать к разработке подходящего вопросника(ов). Вопросник играет центральную роль в той части процесса обследования, в которой информация передается от тех, кто ею располагает (респондентов), тем, кому она необходима (пользователям). Это — инструмент, с помощью которого информационные потребности пользователей выражаются в эксплуатационных терминах и который является основным источником ввода информации в систему обработки данных обследования.

20. Согласно работе Ланделла (2003 год), физическая компоновка вопросника, который предполагается использовать на этапе регистрации, влияет на процесс получения данных, и наоборот. Если для получения данных наиболее приемлемыми будут определены технологии сканирования, то потребуются специальные формуляры, причем они могут быть различными в зависимости от того, будут ли использоваться методы ввода данных «с клавиатуры на диск» или вручную.

21. Вне зависимости от технологий ввода данных, каждый вопросник должен быть однозначным образом идентифицируемым. Уникальный идентификатор должен быть пропечатан на каждом формуляре. Поскольку неправильное распознавание идентификатора формуляра может привести к дублированию вводимых данных и к другим проблемам, необходимо принять меры для минимизации такого риска. Очевидно, что наилучшим выбором при использовании методов сканирования являются штриховые коды. При рассмотрении ручного варианта ввода данных идентификатор формуляра должен, тем не менее, содержать информацию в виде контрольной цифры для предотвращения неправильного ввода данных. Идентификационный код должен однозначно определять каждый вопросник и при этом всегда должен быть цифровым. Обычно в идентификатор формуляра добавляется также информация о придании весов или факторов расширения выборки (страты, первичные единицы выборки, территориальные сегменты, отличительные признаки административных районов, необходимые для составления таблиц и т. д.).

22. Формуляры обычно брошюруются в виде книг (например, книга для каждого счетного участка или городского микрорайона и т. д.). Каждая книга формуляров, как и сами формуляры, должна быть однозначно идентифицируемой, при этом должна прослеживаться четкая взаимосвязь между книгой и содержащимися в ней формулярами для обеспечения того, чтобы формуляр X всегда относился исключительно к книге Y и только к книге Y. Идентификационная информация книги должна использоваться в течение всего процесса обработки данных, начиная

с доставки книги на склад с места проведения обследования и вплоть до извлечения того или иного формуляра из книги, например, при необходимости какой-либо проверки в ходе составления таблиц и анализа данных. Вследствие этого должен быть сведен к минимуму риск неправильной идентификации конкретной книги, как и в случае с неправильной идентификацией конкретного формуляра.

23. Каждое поле формуляра должно быть предназначено для размещения в нем максимально возможного числа знаков для соответствующей переменной величины; например, максимально возможное число членов домохозяйств должно указываться абсолютно четко, для того чтобы определять правильный размер поля.

24. Важно обеспечить отсутствие ошибок в определении единиц наблюдения, схем пропусков и других аспектов вопросника. В каждом обследовании домашних хозяйств собирается информация по основной статистической единице (базовому объекту) — домашнему хозяйству, а также по различным второстепенным единицам (дополнительным объектам) в рамках домашнего хозяйства — отдельным лицам, статьям бюджета, сельскохозяйственным участкам, возделываемым культурам и т. д. В вопроснике должно быть четко и ясно отражено, что представляют собой включенные в него единицы, при этом необходимо также обеспечить надлежащее обозначение каждой наблюдаемой единицы с помощью уникального идентификационного кода. Типовым методом идентификации домохозяйств, являющимся также важным элементом системы ручного ввода данных, является использование простого серийного номера, который пишется или штампуется на титульном листе вопросника или печатается на нем типографским способом. Обычно серийный номер выступает также в качестве идентификатора формуляра.

25. В условиях, когда все более широкое распространение в качестве способа быстрого получения данных приобретают сканирование и обработка изображений, представляется важным также обсудить некоторые характерные особенности, относящиеся к разработке вопросников для сканирования изображений. Ниже рассматриваются некоторые аспекты структуры вопросника, которые возникают в случаях его возможной обработки сканером изображений, например с помощью программ оптического распознавания символов (ОРС), интеллектуального распознавания символов (ИРС) или оптического считывания меток (ОСМ).

26. Вопросник проверяется по двум штриховым кодам. Первый — это код, идентифицирующий каждую страницу вопросника, что очень важно в тех случаях, когда страницы очень похожи друг на друга по своему внешнему виду и формату. Это основное средство, с помощью которого программа обработки изображений различает отдельные страницы вопросника. Второй штриховой код, обычно с соответствующей интерпретацией, помещается на каждой странице вопросника в абсолютно одинаковом виде, однако отличается согласно определенной последовательности от штрихового кода на каждом последующем вопроснике. Этот штриховой код увязывает между собой страницы определенного вопросника, и это чрезвычайно важно, поскольку при подготовке к сканированию страницы обычно разделяются.

27. Точное расположение полей на распечатанном формуляре по своему внешнему виду представляет собой основной формат данных, которые предполагается получить из конкретного вопросника. Если поставлена задача собирать данные, например, со счетных участков с пятизначным кодом, то в печатном формате вопросника создается поле, разделенное на пять разрядов, и пятизначный код счетного участка должен заноситься в это поле. Если поле сбора информации предназначено для записи данных вручную, то такая точность при печатании вопросника не нужна, и все что требуется, — это наличие открытого поля, достаточного по размеру для записи пяти знаков. Важно, однако, настроить схему системы ввода данных таким образом, чтобы поля

выводились на дисплей в цифровой последовательности и автоматически заполнялись во избежание нарушения расположения знаков и неправильной интерпретации данных.

28. Разработка формуляров для ввода данных путем сканирования также серьезно отличается от впечатывания с клавиатуры, поскольку сканнер для идентификации полностью опирается на расположение полей данных. В противоположность вводу данных вручную получение данных путем сканирования не требует напечатанных обозначений полей на вопроснике, за исключением некоторых корректировочных полей на каждой странице. Для обеспечения высокого уровня точности при распознавании изображений системой сканирования поля не должны быть слишком маленькими или большими. Регистраторов в ходе подготовки и при ведении работ на местах необходимо инструктировать таким образом, чтобы они вписывали четкие и ясные знаки по центру поля данных.

29. Существуют также фундаментальные различия в отношении предусмотренного в вопроснике способа сбора данных, например кодов профессий и занятий, между процессами кодирования вручную, подходящими для ручного ввода данных с помощью клавиатуры, и кодирования на экране, используя выведенную на него автоматизированную справочную таблицу со списком кодов профессий и занятий. Когда вопросник предназначается для ввода данных с клавиатуры, таблица кодов печатается на самом вопроснике для использования сотрудником, отвечающим за внесение кодов в вопросник после рассмотрения ответа на открытый вопрос: «*Ваш род занятий?*». В случае сканирования вопросника резервирование места для такой таблицы на распечатанном вопроснике не является абсолютно обязательным, поскольку справочник кодов встроен в матрицу печати, создаваемую на компьютере, при этом роли специалиста-кодировщика и сотрудника по вводу данных сводятся к проверке достоверности данных. В процессе ввода данных касательно кодов профессий и занятий проверяющему сотруднику предоставляется открытый ответ на основе отсканированного изображения, и для быстрого выбора появляется ниспадающее меню с проиндексированными кодами профессий и занятий.

30. Когда рассматривается возможность использования методов сканирования, важным вопросом становится также качество печати вопросника. Сканеры более чувствительны к дефектам печати, чем человеческий глаз. Проблемы могут возникнуть, например, при использовании некоторых цветов или комбинаций цветов, вариаций цвета и четкости, при наличии перекошенных или сдвинутых знаков, ошибок в автоматической нумерации страниц и ошибок брошюровки.

31. В работах Ст. Катерин (2003 год) и Ланделла (2003 год) рассмотрены некоторые ключевые проблемы, связанные с использованием метода сканирования для обработки результатов статистических обследований и переписей.

9.3.3. Проектирование систем обработки данных в обследованиях домашних хозяйств

9.3.3.1. Обобщенный подход к обработке данных в обследованиях домашних хозяйств

32. Системное проектирование является одним из основных направлений работы, проводимой в рамках планирования обследования домашних хозяйств. По сути, на этом этапе предполагаемые к сбору данные обследования, а также вся система обработки данных обследования конкретизируются в определенную формализованную схему.

33. Джамбва, Парирењатва и Розен в своей работе (1989 год) обсуждают преимущества, получаемые от введения и использования формализованной схемы проектирования, разработки и

документирования всех систем в рамках того или иного статистического агентства, особенно для обследований домашних хозяйств, которые сводятся к следующему:

a) данная схема может стать общей платформой для сотрудничества, которое необходимо между статистиками, профильными экспертами и системными аналитиками/программистами;

b) все операции обследования должны быть четко описаны и оформлены в форме документов, с которыми можно было бы сверяться на более поздних этапах. Итоговая документация (иными словами, набор метаданных) важна как для разработки, так и для обслуживания соответствующих систем составления статистики;

c) расходы статистических агентств на разработку и обслуживание таких систем, как правило, бывают довольно высокими с учетом того, что обычно требуется множество различных систем. Структура обследований домашних хозяйств имеет тенденцию следовать одним и тем же закономерностям и принципам. Например, в них используются одни и те же структуры файлов и данных, системы кодирования и т. д. Поэтому в последующих обследованиях могут успешно использоваться системы обработки данных, разработанные для предшествующих обследований, и это, как ожидается, может привести к снижению расходов на их разработку и обслуживание;

d) использование формализованного подхода также играет важную роль в интегрировании обследований, если, например, есть намерение провести любой комбинированный анализ данных из различных обследований или из различных раундов одного обследования.

9.3.3.2. Общие особенности формализованной схемы системного проектирования

34. В данном разделе указаны некоторые общие и основополагающие аспекты упомянутой выше формализованной схемы.

Структура данных

35. Решения по поводу планируемой к анализу социальной проблемы, планируемых к использованию данных и планируемых к применению статистических методов играют основополагающую роль для качества анализа. Тем не менее еще более важным вопросом является выявление и определение объектов или единиц анализа в ходе обследования. Важность этого аспекта уже подчеркивалась в разделе по разработке и печатанию формуляров. На этапе проектирования системы обработки данных обследования необходимо выполнить более формализованное и подробное описание единицы (объекта) анализа и переменных показателей.

36. Согласно работе Сандгрена (1986 год), объект или единица анализа может определяться как любая конкретная или абстрактная сущность (физический объект, живое существо, организация, событие и т. д.), о котором пользователи данных могут пожелать получить информацию. Определение объекта теснейшим образом увязано с той социальной проблемой (целями обследования), в отношении которой собираются и анализируются данные. Аналогичным образом, применительно к обследованиям домашних хозяйств объектами, предметами, элементами или единицами, по которым заинтересованные стороны хотели бы получить данные, являются, например, домашнее хозяйство, человек, земельные участки и т. д. В большинстве случаев базовым объектом выступает домохозяйство, и обычно существует целый ряд смежных объектов, связанных с базовым объектом, которые зависят от конкретного обследования.

Таблица 9.1
Пример целей/аналитических единиц обследования на основе Зимбабвийского межпереписного демографического обследования 1987 года

Объект /единица	Идентификаторы переменных показателей	Определение объекта/единицы	Важные переменные показатели	Смежные объекты	
				Объект	Внешний ключ
Домохозяйство	HID = идентификатор домохозяйства (территория, административное деление, подразделение, EANR = номер *СУ*, HHNR = номер главы домохозяйства)	Дом — это группа лиц, которые обычно проживают и принимают пищу совместно, за исключением гостей	SOH = размер домохозяйства СТРАТА РАЙОН	Человек Умершее лицо	HID HID
Человек	HID, PID PID = идентификатор лица	Лицо, являющееся обычным членом домохозяйства или гостем, находившимся в нем прошлым вечером	SEX, AGE, MARSTAT = семейное положение, ETHNIC= этническая группа, USMEM = обычный член домохозяйства, RELTH = родственная связь с главой домохозяйства	Домохозяйство Женщина от 12 лет и старше	HID HID, PID
Умершее лицо	HID, DID = идентификатор умершего лица	Умершее лицо, являвшееся обычным членом домохозяйства в течение последних 12 месяцев	SEXD = пол умершего лица AGED = возраст умершего лица	Домохозяйство	HID
Женщина от 12 лет и старше	HID, DID	Каждая женщина в возрасте 12 лет и старше, являющаяся обычным членом домохозяйства или гостем, находившимся в нем прошлым вечером	Число рожденных детей	Человек	HID, DID

37. В таблице 9.1 приводятся объекты/единицы, определенные для Зимбабвийского межпереписного демографического обследования 1987 года (ЗМДО). Домохозяйство было базовым объектом, а смежными объектами — «человек», «женщина от 12 лет и старше» и «умершее лицо» (Лагерлоф, 1988 год).

38. По каждому объекту имеется несколько представляющих интерес переменных показателей. Переменные показатели — это свойства (признаки или характеристики) объектов. Например,

объект «человек» в качестве переменного показателя имеет возраст, доход, род занятий, семейное положение и т. д. Переменные показатели могут быть количественными или качественными.

39. Каждый объект должен также иметь уникальный идентификатор. Идентификатор смежного объекта показывает тот базовый объект, с которым он связан. Например, «человек» связан с «домохозяйством» и идентифицируется комбинацией идентификатора домохозяйства (HID) и идентификатора человека (PID), а именно, порядковым номером по перечню домохозяйств (PID).

Вводные данные для системы обработки данных

40. Вводные данные состоят из величин, полученных и записанных регистраторами, в соответствии с вопросником обследования.

Выходные данные системы обработки данных

41. Выходные данные системы состоят в основном из статистических таблиц (в соответствии с планом составления таблиц), баз данных, содержащих микро- и макроданные, и т. д. Они варьируются в зависимости от типа объекта, типа переменного показателя и вида статистического измерения. Сведенные в таблицу переменные показатели обычно являются «первичными», однако их также можно получать из первичных переменных показателей.

Организация файлов

42. Обычно надлежит иметь различные структуры файлов на этапе ввода и на этапе перед составлением таблиц. Например, при вводе данных для обследований домашних хозяйств может быть отдано предпочтение файлу переменной длины (в отличие от двумерного файла), поскольку домохозяйства отличаются по размеру и составу, отсюда во время ввода данных и возникает необходимость в записях переменной длины. Этот метод эффективно использует пространство, но он неудобен для последующей обработки. В конечном счете, однако, зачастую рекомендуется сводить данные в двумерные файлы для облегчения составления таблиц и оптимального использования различных видов универсального программного обеспечения.

43. *Функциональная схема системы*. Для обследования домашних хозяйств необходимо составить функциональную схему с разумным уровнем детализации. Эта схема важна по многим причинам. Прежде всего, она является инструментом для формирования временных графиков и оценки ресурсов, необходимых для завершения обработки данных обследования. Как правило, основными видами деятельности при обработке данных обследования являются:

a) проверка, редактирование и кодирование данных;

b) ввод, проверка подлинности и правильности данных;

c) преобразование структуры данных, используемой на этапе ввода, в структуру, пригодную для составления таблиц;

d) составление таблиц.

44. Функциональная схема системы должна также включать в себя базовые операции с файлами, такие как отбор, проекция, сортировка и сопоставление файлов, вывод новых переменных, агрегирование, табулирование и графическое изображение.

Система документации

45. Всеобъемлющая и четкая документация (то есть набор метаданных) важна как для разработки, так и для обслуживания систем обработки данных по результатам обследований домашних хозяйств. В связи с этим важно документально оформлять файлы и различные операции, с тем чтобы их также могли использовать люди, которые не участвовали во внедрении первоначальной системы. В целях обеспечения достаточности документации необходимо использовать стандартизованный шаблон и хранить его в электронном виде вместе с данными.

46. По мере возможности, для переменных показателей в системах обработки данных для самых различных обследований организация, проводящая обследование, должна использовать одни и те же названия, коды и формат данных, если коды имеют одинаковое значение. Это особенно важно для тех переменных показателей, которые используются для обозначения записей (объектов) в рамках одного файла, поскольку такие переменные показатели могут также применяться при комбинировании (объединении) данных из различных систем.

47. Для проектировании экранных изображений ввода данных или сканирования формуляров в программные пакеты встроены функции с использованием шаблонов для помощи в составлении документации; и для получения эффективно разработанной документации ими необходимо пользоваться в полном объеме. Например, форматы словаря данных Системы обработки данных переписей населения и обследований (CSPro) или Интегрированной системы обработки данных для микрокомпьютеров (ИСОДМ) определяют позицию каждого переменного показателя в файле данных: начальные и конечные точки, цифровую или буквенную форму переменного показателя, повторяется он или нет, и если да — то сколько раз. Словарь также маркирует величины, содержащиеся внутри переменного показателя (St. Catherine, 2003) (более подробную информацию о ИСОДМ и CSPro см. в добавлении к данной главе).

9.4. Проводимые в рамках обследования операции и обработка данных

9.4.1. Создание инструментария и план выборки

48. Как рассматривалось в главах 3 и 4, единицами выборки первого этапа для многих обследований домашних хозяйств являются счетные участки (СУ) переписи, определенные для последней проведенной национальной переписи. Создание компьютерного файла со списком всех СУ страны — это удобный и эффективный способ формирования инструментария для первого этапа выборки, причем наилучшим образом это достигается с помощью программы обработки крупноформатных таблиц, такой как Microsoft Excel, используя одну строку для каждого СУ и столбцы для всей информации, которая может потребоваться (более подробную информацию по программе Excel см. в добавлении к данной главе).

49. Этот инструментарий должен быть легко доступным и должен использоваться для различных манипуляций, таких как сортировка, фильтрация и составление кратких статистических данных, которые могут потребоваться при планировании выборки и соответствующих расчетах. Программу Microsoft Excel легко применять, и многие знают способы ее использования; она обладает функциями сортировки, фильтрации и объединения, которые необходимы при формировании выборок на основе инструментария. Рабочие таблицы могут легко вставляться в большинство других программных пакетов. Как правило, удобнее всего составлять отдельную рабочую таблицу по каждой страте выборки.

50. Содержание записей по единицам инструментария должно быть следующим:

- Необходимо включить первичный идентификатор, который должен быть числовым. Он должен содержать код, который однозначно идентифицирует все административные районы и подразделения, где находится данная единица инструментария. Лучше нумеровать единицы инструментария в географической последовательности. Обычно коды *СУ* уже имеют указанные выше свойства.

- Важную роль играет также вторичный идентификатор, т. е. название деревни (или иного административного подразделения), где находится данная единица инструментария. Вторичные идентификаторы используются для локализации единицы инструментария на картах и на местности.

- В файл необходимо включать целый ряд характеристик единиц выборки, таких как показатель размера (население, домохозяйства), сельские или городские районы, плотность населения и т. д. Такие характеристики могут использоваться в целях стратификации или установления вероятностей отбора, а также в качестве вспомогательных переменных в расчетах.

- Должны включаться также оперативные данные, такие как информация об изменениях в единицах, а также указание на способ применения выборки.

51. Должны в полном объеме документироваться процедуры отбора, а также показатели вероятности отбора для всех единиц выборки на каждом ее этапе. При применении эталонных выборок должны иметься записи о том, какие единицы эталонной выборки были использованы в выборках для конкретных обследований. Для единиц выборки следует использовать стандартную систему идентификационных номеров.

52. В качестве примера рассмотренных выше тезисов может служить эталонная выборка в Намибии, базирующаяся на данных по переписи населения и жилого фонда 1991 года (Центральное статистическое бюро Намибии, 1996 год).

Пример

Для возможности отбора случайной выборки географических районов в Намибии возникла необходимость в формировании инструментария выборки географических районов. Для этой цели был создан инструментарий географических районов — первичных единиц выборки (*ПЕВ*). В среднем в этих районах находилось примерно по 100 домохозяйств, при этом большинство имело в своем составе от 80 до 100 домохозяйств. Эти районы были сформированы на основе счетных участков переписи населения и жилого фонда 1991 года. Мелкие *СУ* были объединены с прилегающими *СУ* для формирования *ПЕВ* достаточного размера. Использовалось эмпирическое правило, что каждая *ПЕВ* должна содержать не менее 80 домохозяйств. Всего было сформировано около 1 685 таких *ПЕВ*, разбитых в соответствии со стратами *ПЕВ* по регионам, по сельским, полугородским и городским районам.

53. Такая стратификация базировалась на классификации *СУ*, проведенной в ходе подготовки переписи 1991 года. Всего было создано 32 страты, в рамках которых списки *ПЕВ* были сформированы в географической последовательности. В городских и полугородских районах также составлялись списки *ПЕВ* в разбивке по уровню дохода конкретного района. Первыми в списках страт городских и полугородских районов были районы с высоким доходом, затем — со средним/низким доходом.

54. Центральное статистическое бюро подготовило в формате Microsoft Excel файлы инструментария эталонной выборки *ПЕВ*. Этот инструментарий содержал следующие позиции:

- регион
- уникальный номер *ПЕВ*
- уровень дохода (только для городских районов)
- район
- номер(а) *СУ*
- число домохозяйств согласно переписи 1991 года
- суммарное количество домохозяйств по страте
- население в разбивке по половой принадлежности согласно переписи
- статус применительно к эталонной выборке (входила ли данная *ПЕВ* в эталонную выборку или нет)
- номер *ПЕВ* в эталонной выборке (только для *ПЕВ*, входивших в эталонную выборку)
- веса (повышающие или инфляционные коэффициенты) только для *ПЕВ*, входивших в эталонную выборку.

55. Для каждого региона существовал один файл Microsoft Excel, и в рамках каждого такого файла Excel единицы *ПЕВ* были сгруппированы следующим образом: сельские районы; полугородские, городские районы — с высоким доходом; и полугородские, городские районы — со средним/низким доходом.

56. Петтерсон в своей работе (2003 год) детально рассматривает вопросы, связанные с эталонной выборкой. Муньос (2003 год) дополнительно рассматривает, каким образом компьютеризованный инструментарий, подобный описанному выше, может также служить средством для проведения процедуры формирования выборки в целях обследования домашних хозяйств путем управления этой процедурой на ее основных этапах: формирование инструментария первого этапа, обычно основанного на результатах последней переписи населения и жилого фонда (*СУ*); формирование первичных единиц выборки с вероятностью, пропорциональной размеру (размер, измеряемый численностью домохозяйств, жилищных единиц или населения); и обновление крупноформатных таблиц на основе списка отобранных домохозяйств, а также расчет вероятностей отбора и соответствующих им весов выборки. В разделе 9.4.3.5, ниже, по процедурам точечной оценки и расчету весов подробно рассматривается расчет вероятностей отбора и соответствующих им весов (также см. главу 5). Требуемые для таких расчетов данные могут быть получены из крупноформатной таблицы, которая обсуждалась выше, а расчет весов может осуществляться с использованием такой таблицы, как это показано в работе Муньоса.

9.4.2. Сбор данных и управление данными

57. В ходе обследований домашних хозяйств может быть получено большое количество заполненных вопросников. Поэтому во избежание хаоса еще на ранних этапах необходимо тщательно продумать и сформулировать процедуры физической обработки и учета этого массива документов. Порядок ручной обработки (составление папок и извлечение документов из них) вопросников должен быть тщательно спланирован и организован задолго до того, как с мест начнут поступать данные. Одним из важнейших аспектов такой системы является обязательная оценка ожидаемого

объема данных, с тем чтобы можно было приобрести необходимое количество папок, коробок и т. д. и чтобы резервировать место для их хранения на полках или в шкафах. Вторым аспектом такой системы является ведение журнала, в который заносится информация по вопросникам после их получения и по которому можно отслеживать прохождение потока данных через всю систему. Это — ключевые аспекты обслуживания данных и важнейшие предварительные условия для успешного управления и осуществления стратегии обработки данных любого обследования.

58. Физическая сохранность заполненных вопросников также является важной проблемой на этом этапе. Именно в данной сфере особенно ценным оказалось использование сканирования изображений. Если вопросники сканируются сразу после поступления в офис, снижается риск потери данных в этих вопросниках в результате каких-либо возможных инцидентов. Сканирование изображений обеспечивает дополнительный уровень надежности, так как позволяет делать резервные копии вопросников непосредственно в офисе или за его пределами уже после того, как они были отсканированы (Edwin, 2003). Следует, однако, отметить, что успешное внедрение такой технологии в значительной мере зависит от организации и управления процессами ее использования и связанными с этим процедурами в каждом конкретном учреждении. Вследствие этого, хотя сканирование успешно применялось в некоторых странах, оно не принесло успеха в других странах. Для успеха операции сканирования необходимо, среди прочих факторов, учитывать структуру организации статистического бюро с точки зрения централизации или децентрализации соответствующих процедур; квалификацию проводящих обследование сотрудников и гарантию качества инструментов сбора данных обследования.

9.4.3. Подготовка данных

59. Собранные данные необходимо ввести в файл данных. Перенос данных из вопросников в машиночитаемую форму называется вводом данных. В этой связи часто возникает необходимость разбивать по категориям переменные параметры, по которым даются открытые ответы; такой процесс категоризации называется кодированием. Путем редактирования полученных данных можно выявлять ошибочные данные. В этом случае необходимо принять надлежащие меры для проверки предполагаемых ошибок путем, например, повторного контакта с источником информации. За такими проверками может следовать обновление (коррекция) данных. Этапы обработки включают: ввод, кодирование, редактирование, проверку и обновление/коррекцию данных. В совокупности указанные операции называются в этом руководстве этапом подготовки данных для обработки результатов обследования.

9.4.3.1. Стратегии подготовки данных

60. В работе Муньоса (2003 год) подробно рассматриваются различные аспекты и форматы для подготовки данных. Наиболее распространенной в обследованиях домашних хозяйств организационной формой является проведение централизованной подготовки данных в отведенных для этого пунктах после сбора данных на местах. Альтернативной формой является включение операции ввода данных в качестве составной части работ на местах. Одной из последних инноваций в этой области являются методы проведения опросов с помощью компьютера.

Централизованная подготовка данных

61. До появления персональных компьютеров этот вариант был единственно возможным. В значительной мере он остается основным подходом, используемым при проведении обследований

в развивающихся странах, с некоторыми изменениями в связи с внедрением микрокомпьютеров. В рамках данного подхода применение ввода данных рассматривается как технический процесс, осуществляемый в одном или нескольких местах после проведения опросов. Он может осуществляться в центральных учреждениях национальных статистических служб или в их региональных отделениях.

Подготовка данных на местах

62. Недавнее внедрение компьютерных методов контроля качества в проведении работ на местах рассматривается как один из ключевых факторов повышения качества и своевременного выполнения обследований домашних хозяйств. Согласно этой стратегии ввод данных и контроль за их согласованностью проводятся в качестве неотъемлемого элемента операций на местах.

63. Одной из возможных форм такого подхода являются работа оператора по вводу данных с персональным компьютером в каком-либо постоянном месте (например, в региональном отделении национальной статистической службы) и такая организация работ на местах, при которой оставшаяся часть группы как минимум дважды посещает каждый район проведения обследования (как правило, первичную единицу выборки) таким образом, чтобы дать оператору время для ввода и проверки согласованности данных между выездами. В ходе второго и последующих посещений регистраторы повторно задают соответствующим домохозяйствам вопросы, в ответах на которые программой ввода данных были выявлены ошибки, пропуски или несогласованности.

64. Другим подходом является работа оператора по вводу данных с портативным компьютером вместе с остальной группой во время посещений мест проведения обследования. Вся группа остается в том или ином месте до тех пор, пока все данные не будут введены и подтверждены программой ввода данных как полные и точные.

65. Ощутимыми сравнительными преимуществами объединения процессов сбора и подготовки данных считаются: возможность достижения более высокого качества данных, поскольку ошибки могут быть исправлены, пока регистраторы еще находятся на местах; возможность создания баз данных, а также составления таблиц и анализа вскоре после завершения работ на местах; более обширные возможности стандартизации процесса сбора данных регистраторами.

66. В рамках обоих описанных выше подходов критически важным фактором является наличие постоянного электроснабжения в местах проведения операций. В странах с неудовлетворительным электроснабжением, какими являются большинство развивающихся стран, особенно в их сельских районах, такие операции просто неосуществимы. Необходимо также отметить наличие организационных и логистических проблем, связанных с использованием переносного оборудования для сбора и подготовки данных. Предварительными условиями для успешного применения такой стратегии являются наличие эффективной системы управления; действенные меры обеспечения сохранности оборудования и данных; наличие адекватных средств хранения резервных копий данных; достаточное снабжение расходными материалами, такими как запасные батареи, для использования в работах на местах.

Проведение опроса с помощью компьютера

67. Личный опрос с помощью компьютера (ЛОПК) представляет форму проведения личного опроса, при которой регистратор вместо заполнения бумажного вопросника использует переносной или карманный компьютер для введения данных непосредственно в базу данных. Этот метод экономит время в ходе обработки данных и избавляет регистратора от необходимости носить

с собой сотни вопросников. Тем не менее, хотя такая технология существует уже многие годы, очень мало было сделано для серьезного применения этой стратегии для комплексных обследований в развивающихся странах. Такой метод сбора данных может быть весьма дорогостоящим с организационной точки зрения и, кроме того, требует от регистраторов навыков пользования компьютером и печатания с помощью клавиатуры. Опрос с применением компьютера также требует хорошо структурированной организации опросов — от начала и до конца. При этом для большинства обследований в развивающихся странах требуются несколько посещений каждого домохозяйства, отдельные опросы каждого члена домохозяйства и т. д., в процессе, который не является строго структурированным, а по существу определяется самим регистратором.

9.4.3.2. Кодирование и редактирование данных по результатам обследований

68. Проверка, редактирование и кодирование данных представляют собой, пожалуй, наиболее сложную фазу обработки данных. Именно в сфере организации системы управления данными и их подготовки недавно обученные специалисты по проведению обследований зачастую испытывают наибольшие трудности. Если есть такая возможность, то наилучшим образом процедуры проверки, редактирования и кодирования данных осуществляются в автоматизированном режиме. Однако применительно к кодированию требуется, естественно, учитывать случаи, когда коды невозможно назначать автоматически, при этом возникает необходимость в ручном присвоении кодов.

Кодирование

69. Целью этой операции является подготовка данных в той форме, которая применима для их ввода в компьютер. Операция кодирования в основном включает присвоение цифровых кодов ответам, записанным словами (например, касательно географического местоположения, рода деятельности, отрасли и т. д.). Данная операция может также предусматривать перезапись, при которой уже назначенные и зафиксированные в ходе опроса цифровые коды переносятся в кодовые таблицы.

70. Для специалистов-кодировщиков надлежит подготовить пособие с исчерпывающими инструкциями. Такое пособие должно содержать набор не связанных друг с другом категорий, охватывающих все возможные ответы на вопросы в рассматриваемой области. Для крупномасштабного обследования желательно в максимальной степени использовать закрытые или заранее закодированные вопросы.

Редактирование и проверка данных

71. Целями проверки и/или редактирования вопросников являются: *a*) достижение согласованности между данными, а также согласованности внутри таблиц данных и между ними, и *b*) выявление и проверка, коррекция или устранение резко отклоняющихся значений, поскольку предельные значения являются основным источником вариантности выборок в оценках обследований.

72. Редактирование включает пересмотр или коррекцию записей в вопросниках. Оно может рассматриваться как процедура проверки достоверности, с помощью которой выявляются и корректируются несоответствия и невозможные значения данных, или как статистическая процедура, в рамках которой проводятся проверки на основе статистического анализа данных. Существует тенденция во все больших масштабах осуществлять редактирование данных на компьютере либо при вводе данных, либо в ходе специальных редакторских «прогонов» данных. Такие редакторские прогоны могут быть, а могут и не быть интерактивными, иными словами, оператор может либо проводить, либо не проводить немедленную коррекцию ошибок. Тем не менее устранение

более сложных ошибок может потребовать больше времени и более глубокого анализа до того момента, когда будет найден правильный метод коррекции, и именно для этой цели более подходящими являются неинтерактивные редакторские прогоны. В справочном материале, подготовленном в работе Олссона (1990 год), более подробно рассматриваются различные аспекты проверки и редактирования данных по результатам обследований.

73. *Проверка и редактирование вручную.* Основной задачей проверки и редактирования в ручном режиме является выявление пропусков, несоответствий и прочих очевидных ошибок в вопросниках до начала следующих этапов обработки данных. Редактирование вручную следует начинать как можно скорее и как можно ближе к источнику данных, например, на уровне областного, окружного или местного отделения. В идеальном случае большинство ошибок в данных должно быть выявлено и исправлено в ходе работ на местах до направления формуляров в центр обработки данных. В связи с этим в пособиях и инструкциях регистратору и контролеру обычно поручается проверять вопросники и исправлять любые ошибки прямо на местах до отправки данных. Это — важная и сложная задача, от выполнения которой напрямую зависит качество работ на местах, эффективность контроля, управление обследованием и т. д.

74. *Редактирование с помощью компьютера.* Компьютерное редактирование может осуществляться: *a*) в интерактивном режиме на этапе ввода данных, *b*) методами пакетной обработки после ввода данных или *c*) с использованием некоторой комбинации методов *a* и *b*. Интерактивное редактирование обычно представляется более полезным применительно к простым ошибкам (например, опечаткам), при этом оно задерживает процесс ввода данных в случаях, когда те или иные ошибки требуют консультации с контролерами. Обработка таких ошибок, включая ошибки неполучения ответов, должна выделяться в отдельную операцию компьютерного редактирования.

75. Программы компьютерного редактирования зачастую создаются с использованием программ баз данных, таких как Интегрированная система обработки данных для микрокомпьютеров (ИСОДМ), Комплексная система для анализа результатов обследования (ИССА), Система обработки результатов переписей и обследований (CSPro), Visual Basic и Microsoft Access (более подробная информация по указанным программам приводится в добавлении к данной главе). Простейшие программы сканируют данные, запись за записью, и отмечают несогласованности, исходя из занесенных в программу правил редактирования. В более сложных программах редактирования может осуществляться сравнение переменных показателей (например, идентификационных переменных) между файлами и отмечаться расхождения. Результатом работы таких систем являются списки ошибок, которые чаще всего проверяются в ручном режиме в сравнении с необработанными данными. Ошибки исправляются в копии файла необработанных данных.

Виды проверок

76. Занесенные в вопросники данные требуют различных видов проверки, при этом наиболее типичными являются проверка по диапазону значений, сверка с контрольными данными, проверка пропущенных данных, проверка согласованности данных и проверка на наличие опечаток (Munoz, 2003).

77. *Проверка по диапазону значений.* Проверки по диапазону значений предназначены для обеспечения того, чтобы каждый переменный показатель обследования содержал данные, находящиеся в пределах ограниченного диапазона истинных значений. Категорийные переменные могут иметь только одно из заранее определенных для них в вопроснике значений (например, пол может кодироваться исключительно, как «1» для мужчин и «2» для женщин). Хронологические перемен-

ные должны содержать достоверные даты, а числовые переменные должны находиться в пределах заранее определенных минимальных и максимальных значений (например, возраст — от 0 до 95 лет). Особый случай проверки по диапазону значений имеет место, когда появляется возможность сравнения данных из двух или более тесно связанных между собой областей с внешними справочными таблицами.

78. *Проверка пропущенных данных.* В ходе проверок пропущенных данных удостоверяется, были ли надлежащим образом соблюдены схемы и коды пропуска данных. Например, с помощью простой проверки удостоверяется, что вопросы, относящиеся исключительно к категории «школа+дети», не записываются для ребенка, который ответил «нет» на первоначальный вопрос о посещении школы. В ходе более сложной проверки может быть установлено, были ли в отношении каждого респондента заполнены надлежащие модули вопросника. Предполагается, что каждый член домохозяйства, в зависимости от своего возраста и пола, должен ответить на конкретные разделы вопросника (или пропустить их). Например, женщины в возрасте 15–49 лет могут быть включены в раздел по рождаемости, а мужчины — не могут.

79. *Проверка согласованности данных.* С помощью проверок согласованности данных устанавливается, соответствуют ли значения, указанные в ответе на один вопрос, значениям из другого вопроса. Такая проверка будет простой, если оба значения относятся к одной и той же статистической единице, например дата рождения и возраст, указанный тем или иным человеком. Более сложные проверки согласованности данных предусматривают сравнение информации по двум или нескольким различным единицам наблюдения. Например, родители должны быть не менее чем на 15 лет старше своих детей.

80. *Проверка опечаток.* Типичная опечатка заключается в перестановке цифр в числе (например, ввод числа «14» вместо «41»). Такая ошибка в возрасте может быть выявлена проверкой соответствия данных с семейным положением или родственными связями. Например, находящийся в браке или вдовый взрослый человек в возрасте 41 года, ошибочно указанного как 14 лет, будет отмечен флажком ошибки при сверке возраста с семейным положением. При этом, однако, такая же ошибка в ежемесячных расходах на покупку мяса может легко остаться незамеченной, поскольку достоверными значениями могут быть как 14, так и 41 доллар США. Обычный способ решения такой проблемы предусматривает двойное введение данных из каждого вопросника двумя разными операторами.

Решение проблемы отсутствующих данных

81. Когда проведение обследования достигает этапа обработки данных, вне всякого сомнения, будет отмечен значительный объем недостающих данных. Некоторые домохозяйства могут переселиться или отказаться отвечать на вопросы. На некоторые вопросы, содержащиеся в вопроснике, могут быть не получены ответы; или некоторые данные могут оказаться ложными или несоответствующими другой информации в вопроснике. Вне зависимости от причины, результатом остается отсутствующая, пустая или частично пустая запись.

82. Важно проводить различие между отсутствующими данными — иными словами, данными, которые должны присутствовать, но точное значение которых неизвестно — и нулевыми данными. Скажем, один вопросник может оказаться пустым, потому что определенное домохозяйство отказалось участвовать в обследовании, в то время как часть другого вопросника может быть незаполненной, поскольку второе домохозяйство, например, не посеяло на своих полях никаких сельхозкультур. Во втором случае переменная величина «посевная площадь» будет нулевой. Такие записи должны сохраняться в файле для анализа и включения в таблицы.

83. Подход, применяемый в отношении действительно отсутствующих данных, зависит от того, какие именно данные отсутствуют. Отдельный элемент выборки может полностью отсутствовать по причине отказа домохозяйства от участия в обследовании или неспособности респондента данного домохозяйства ответить на весь набор вопросов в вопроснике. Такие случаи обозначаются, как «неполучение ответов от единицы».

84. Если респондент способен ответить только на некоторые вопросы, имеет место случай «неполучения ответов на отдельные вопросы», поскольку от рассматриваемого домохозяйства были получены некоторые, но не все данные.

85. Отсутствующие данные любого типа приводят к появлению искаженных оценок обследования, как постоянно подчеркивается в настоящем руководстве. Более подробную информацию о надлежащей обработке случаев неполучения ответов, включая методы внесения поправок на такие случаи, см. в главе 6.

86. В случае неполучения ответов на отдельные вопросы в целях обеспечения согласованности суммарных значений может возникнуть необходимость замещения отсутствующих данных некоторыми приемлемыми оценками. Этот метод известен под названием «вменение значений», как отмечено в главе 6. Существуют несколько подходов, которые можно использовать для вменения замещаемых значений. Ниже приводятся некоторые из них:

- *Вменение среднего значения:* вменение отсутствующего значения, используя среднюю величину (в рамках *ПЕВ* или всего набора данных);
- *Вменение значений с подбором из недавно собранных данных:* заимствование отсутствующих значений из («донорской») записи, аналогичной незавершенной записи. Такая донорская запись должна пройти все проверки на этапе редактирования;
- *Статистическое вменение:* использование связи (регрессии, соотношения) с какой-то другой переменной, полученной из полного набора данных, для вменения отсутствующего значения.

87. Выше перечислены лишь некоторые из методов вменения значений: для этой цели существует еще несколько методов. Эффективность процедуры вменения, безусловно, зависит от того, насколько успешно модель вменения улавливает суть случая неполучения ответов. В плане выбора имеющейся вспомогательной информации важно, чтобы переменная величина коррелировалась с подлежащей вменению величиной (более подробную информацию по этому вопросу см. в работе Олссона (1990 год)).

9.4.3.3. Ввод данных

88. Целью ввода данных является перевод информации с выполненных на бумаге вопросников в промежуточный продукт (машиносчитываемые файлы), который требует дальнейшей обработки с помощью программ редактирования и технических процедур для получения так называемой чистой базы данных в качестве конечного продукта. На начальной фазе ввода данных приоритетами являются скорость операций и обеспечение того, чтобы информация в файлах полностью соответствовала собранной информации, содержащейся в вопросниках.

89. Решение о методе ввода данных из вопросников на компьютерные носители должно приниматься на самом раннем этапе, поскольку это оказывает существенное влияние на основную организацию рабочего процесса, технологию хранения данных, макет формуляров, а также кадровую структуру.

Ввод данных с клавиатуры на диск

90. Ввод данных с клавиатуры на диск предусматривает перенос с помощью клавиатуры закодированных данных, например, на диск, дискету или компакт-диск. Многие проводящие обследования организации в развивающихся странах накопили значительный опыт в применении этого метода ввода данных. Это основной применяемый подход, и его использование было подкреплено распространением персональных компьютеров и соответствующего программного обеспечения.

91. *Программа ввода данных.* Обычно программа ввода данных включает три модуля. Первый — это модуль ввода всей информации, второй — модуль проверки достоверности введенных данных, с помощью которого подтверждается надлежащее качество введенной информации и ведется наблюдение за работой операторов по вводу данных. Третий — модуль коррекции введенной информации, поскольку может возникнуть необходимость исправления ошибок в оценках, не выявленных во время процессов ввода данных или проверки их достоверности.

92. Обычно программа ввода данных имеет основное меню, по которому сотрудник, отвечающий за ввод данных, может выбирать между режимами ввода, проверки и коррекции данных. До работы в основном меню пользователь должен с помощью имени пользователя и пароля подтвердить наличие у него/нее разрешения на вход в программу. Если логин не будет принят (иными словами, при вводе неправильного имени или пароля), программа немедленно закроется. Все имена и пароли пользователей хранятся в таблице пользователей в вычислительной машине базы данных, где зашифрованы пароли. Когда пользователь заходит в систему с действующим паролем, таблицы в вычислительной машине базы данных обновляются.

93. *Модуль ввода данных.* Модуль ввода данных — это связующее звено между вопросником и файлом данных или базой данных. Система ввода должна быть очень простой в использовании для оператора по вводу данных. Существует ряд важных требований, предусматривающих, что:

- Экран ввода данных должен выглядеть максимально похожим на соответствующие страницы вопросника. Оператор должен иметь возможность очень быстро находить на экране соответствующий раздел вопросника.
- Чрезвычайно важна скорость ввода данных. Оператор не должен ожидать, пока система оценит каждый вводимый показатель. Поэтому требуется, чтобы процесс оценки был очень быстрым; это предполагает, что система не должна контактировать с сервером больше, чем необходимо, в нашем случае это означает, что показатели не будут сохраняться в базе данных до тех пор, пока не будут введены все показатели по конкретному домохозяйству. Недостатком здесь является то, что информация, вводимая в данный момент по тому или иному домохозяйству, будет потеряна, если по какой-либо причине программа будет закрыта. Тем не менее более важно обеспечить преимущество относительного увеличения скорости.
- Каждый показатель вопросника должен иметь цифровой код, чтобы обеспечить возможность использования цифровой клавиатуры, что обусловит высокую скорость.
- Модуль ввода данных должен иметь функцию текущего контроля достоверности вводимой величины, когда оператор немедленно получает уведомление об ошибке при введении недействительной величины. Контроль достоверности должен также обслуживать связанные друг с другом показатели: например, если показатель пола имеет значение "1" (мужской), тогда блок информации о рождаемости должен отключаться.
- Программа ввода данных должна, конечно, отмечать как ошибки любые логически невозможные ситуации (как, например, дочь, имеющая возраст старше своей матери)

или маловероятные ситуации (как, например, дочь, имеющая возраст менее чем на 15 лет моложе своей матери).

- Важно отслеживать скорость печатания по количеству ударов по клавишам и время ввода данных для последующего статистического использования, например для прогнозирования суммарного времени ввода данных.

94. *Модуль проверки достоверности данных*. Целью системы проверки достоверности данных является предоставление информации о качестве введенных данных и долю ошибок для каждого оператора по вводу данных. Экран этого модуля должен быть абсолютно идентичным с экраном ввода данных без каких-либо видимых различий. При этом основным отличием является не только определение скорости печатания исходя из числа ударов по клавишам, но и выявление числа ошибок. Функция выбора типа проверки включает *общую проверку*, при которой проверяется достоверность всех *СУ* и вопросников по каждому *СУ*, или *выборочную проверку*, когда проверяются лишь некоторые *СУ* и некоторые вопросники.

95. *Модуль коррекции данных*. Модуль коррекции данных используется в основном для коррекции информации, которую, по той или иной причине, невозможно было завершить в модуле ввода данных. В этом модуле есть возможность добавлять, исключать или обновлять информацию, начиная с полной информации о домохозяйстве и кончая отдельным показателем.

96. *Программа управления работой контролеров.* Программа управления — это инструмент, с помощью которого контролеры вносят изменения в базу данных. Этот инструмент используется главным образом для коррекции пакетного мастер-файла (ПМФ) и для получения отчетов о показателях работы пользователей. Важно, чтобы:

- контролеры имели полный контроль над ПМФ исходя из этой программы; у них должна быть возможность добавлять, удалять и обновлять информацию в ПМФ;
- была возможность добавления и удаления пользователей и получения полного списка всех пользователей; должна быть возможность проверки текущего статуса всех пользователей или только одного пользователя;
- была возможность просмотра и распечатки статистики скорости в отношении печатания; была возможность выбора различных периодов времени;
- была возможность просмотра и распечатки процента ошибок для отдельного пользователя и среднего процента ошибок по всем пользователям;
- была возможность перезагрузки того или иного *СУ* для ввода или проверки достоверности данных;
- была возможность получения из этой программы всей информации, необходимой контролерам для управления работой.

Различные аспекты систем ввода данных с клавиатуры на диск подробно рассматриваются в работе Свенссона (1996 год).

97. *Платформы систем ввода данных с клавиатуры на диск*. На рынке представлено множество платформ для разработки программ ввода и редактирования данных. Например, Система обработки данных переписей населения и обследований, а также ее предшественница — Интегрированная система обработки данных для микрокомпьютеров — доказали свои возможности обеспечивать разработку эффективных программ ввода и редактирования данных для комплексных национальных обследований домашних хозяйств во многих развивающихся странах. К тому же оказалось, что эти платформы очень просто приобрести и что они просты в использовании (Munoz, 2003).

Сканирование

98. Быстрыми темпами растет применение сканирования в обработке данных переписей и обследований. Всего лишь несколько лет назад ввод данных ассоциировался прежде всего с системами, работающими с помощью клавиатур. Многочисленные конкурирующие системы еще не вышли на рынок. Сегодня ситуация изменилась, и все «бестселлеры» из числа систем ввода данных базируются на методах сканирования. Существует несколько вариантов этих методов, причем каждый из них имеет свои достоинства и недостатки. Наиболее широко используемыми методами являются: оптическое распознавание символов (ОРС), означающее распознавание напечатанных знаков; интеллектуальное распознавание символов (ИРС), означающее распознавание написанных от руки знаков; оптическое считывание меток (ОСМ), означающее распознавание сделанных ручкой или карандашом меток, нанесенных в определенных местах — обычно в соответствующих ячейках; считывание штрих-кодов (СШК), означающее распознавание данных, зашифрованных в типографским способом нанесенных штриховых кодах.

99. Согласно работе Ланделла (2003 год), в плане использования технологий сканирования для статистических обследований и переписей выбор обычно делается между интеллектуальным распознаванием символов и оптическим считыванием меток. Страны с большой численностью населения делают выбор в пользу оптического считывания меток, в то время как для комплексных вопросников предпочтение отдается интеллектуальному распознаванию символов. Оптическое считывание меток ограничивает формат вопросника, однако предлагает быструю обработку данных и требует сравнительно менее квалифицированного персонала. Интеллектуальное распознавание символов дает свободу в выборе формата вопросника, однако обработка данных налагает более серьезные требования в плане мощности компьютеров и квалификации персонала. Штрих-коды обычно используются только для введения и извлечения идентификационной информации, например номеров формуляров, поскольку штрих-код содержит информацию в виде контрольного знака для минимизации ошибок.

100. В ходе процесса сканирования вопросники сканируются со скоростью от 40 до 90 двусторонних листов в минуту. Скорость выступает наиболее важным определяющим фактором при предпочтении сканирования над традиционными формами ввода данных с клавиатуры. Программы сканирования в этом случае используются для идентификации страниц вопросника и оценки их содержания с применением интеллектуального распознавания символов и оптического считывания меток. Позиции, по которым возникли вопросы или которые требуют кодирования, направляются проверяющему сотруднику, который просматривает нечетко написанные пункты и кодирует открытые вопросники в соответствии с встроенными в шаблон сканирующей программы справочными электронными таблицами. Присутствует большая доля гибкости в плане способов проведения проверок достоверности данных в зависимости от установленного формата шаблона сканирующей программы. Важнейшие показатели могут просматриваться частично или полностью для максимизации точности ответа, заносимого в файл данных.

101. Было доказано, что использование процессов сканирования изображений может повысить эффективность ввода данных на 70 процентов (Edwin, 2003). Многие проблемы, связанные со сканированием, можно преодолеть с помощью надлежащей технической организации этого процесса. Например, проблема отсутствующих или неправильно расположенных страниц может решаться путем использования штрих-кодов, заранее напечатанных на вопросниках и применяемых в качестве средства увязки между собой различных страниц вопросника. При надлежащем обслуживании и контроле за работой оборудования и программного обеспечения расходы на операцию сканирования (включая закупку оборудования и программного обеспечения) в долгосрочном плане могут оказаться значительно меньшими, чем при использовании клавиатуры для ввода данных.

102. Пока опыт применения сканирования для обследований домашних хозяйств весьма ограничен, особенно в регионе к югу от Сахары. Тем не менее данный метод достаточно широко использовался в ходе раунда переписей населения и жилого фонда 2000 года, и, по всей вероятности, это знаменует собой поворотный пункт в его широком внедрении. Например, в ходе последних переписей сканирование использовалось в Кении, Объединенной Республике Танзании, Южной Африке, Намибии и Замбии. Недавно оно также использовалось в организованных Всемирным банком обследованиях с применением Вопросника по основным показателям благосостояния. Кроме того, такие страны, как Намибия и Южная Африка, приняли эту технологию для своих программ обследований домашних хозяйств.

9.4.3.4. Структура файла и организация наборов данных

Хранение данных

103. Для результатов обследований домашних хозяйств, которые обычно содержат информацию как на уровне домохозяйств, так и на уровне отдельного лица, эффективное использование объема памяти может означать применение последовательных форматов файлов или форматов файлов переменной длины, поскольку различные домохозяйства включают разное число проживающих в них лиц. Двумерный файл, который занимает неоправданно большое место в памяти, был бы применим только в том случае, если бы все вопросы касались домохозяйства как статистической единицы, однако, как указывалось выше, это далеко не так. Некоторые вопросы относятся к второстепенным статистическим единицам, обозначаемым переменными величинами в каждом домохозяйстве, таким как люди, урожай, потребляемые товары и т. д. Хранение данных о возрасте и поле каждого члена домохозяйства в качестве различных переменных величин на уровне домохозяйства было бы неэкономным, поскольку число необходимых переменных величин будет определяться размером самого крупного, а не среднего домохозяйства.

104. Обычно для ввода данных в обследованиях домашних хозяйств используется файл переменной длины. Поскольку домохозяйства различаются по размеру и составу, в процессе ввода данных возникает необходимость в записях переменной длины. Хотя каждый вид записей будет фиксированным по длине и формату, в пределах одного файла будут включаться различные виды записей. Каждый файл будет, по сути, компьютерным изображением заполненного вопросника. Каждая строка или раздел вопросника будет представлять собой ту или иную запись. Каждая запись будет начинаться с последовательности идентификаторов, увязывающих запись с домохозяйством, единицей наблюдения и т. д. Этот метод эффективно использует объем памяти, однако является неудобным для дальнейшей обработки, в ходе которой крайне важное значение приобретают перекрестные ссылки на данные из различных файлов.

105. В пакете CSPro, например, используется структура файлов, которая прекрасно справляется с проблемами, возникающими при работе с различными статистическими единицами, сводя при этом к минимуму требования к объему памяти и взаимодействуя со статистическими программами на этапе анализа.

106. Структура данных сохраняет взаимное однозначное соответствие между каждой наблюдаемой статистической единицей и записями в компьютерных файлах, используя различные типы записей для каждого вида статистической единицы. Например, для управления данными, указанными в перечне домохозяйства, тип записи будет определяться для переменных показателей данного списка, и данные для каждого отдельного лица будут сохраняться в отдельной записи этого типа. Аналогичным образом, в разделе потребления продуктов питания тип записи будет

соответствовать конкретным продуктам питания, и данные, соответствующие каждой отдельной позиции, будут сохраняться в отдельных записях этого типа.

107. Допускается варьирование числа записей каждого типа. Это экономит требуемый объем памяти, поскольку нет необходимости, чтобы файлы предусматривали максимально возможный размер для каждой записи.

108. После включения идентификаторов фактические данные, регистрируемые в рамках обследования по каждой конкретной единице, записываются в полях фиксированной длины в той же последовательности, что и вопросы вопросника. Все данные хранятся в стандартном формате ASCII (Американский стандартный код для обмена информацией).

109. Более подробно вопросы управления файлами в обследованиях домашних хозяйств рассматриваются в работе Муньоса (2003 год) и публикации Всемирного банка (1991 год).

Реструктуризация наборов данных для дальнейших операций

110. Для обеспечения надлежащего анализа соответствующая база данных должна содержать всю информацию о процедуре формирования выборки: маркировки страт плана выборки, первичные единицы выборки, вторичные единицы выборки и т. д., а также веса выборки для каждой единицы выборки. Эта информация понадобится для расчета требуемых статистических показателей и для расчета ошибок выборки для этих показателей.

111. После ввода данных часто возникает необходимость в реструктуризации набора данных и формировании новых файлов, а также в перекодировании некоторых из имеющихся полей данных, с тем чтобы определить новые переменные показатели — более удобные для табуляции и анализа. Это может понадобиться для обеспечения проведения определенных операций с данными, включая процедуру расчета.

112. Первоначальный полный файл данных обследования на практике может содержать информацию по единицам, включенным в выборку из различных совокупностей (Rosen, 1991). Например, применительно к обследованию бюджета домохозяйств в одном и том же первоначальном файле могут содержаться данные по вошедшим в выборку домохозяйствам, а также по вошедшим в выборку лицам. Для оценки статистических характеристик совокупности домохозяйств и совокупности лиц необходим файл с одной записью для каждого вошедшего в выборку домохозяйства и файл с одной записью для каждого вошедшего в выборку лица, соответственно. Наборы данных или файлы, базирующиеся на домохозяйствах как на единицах (объектах), используются для составления статистики (таблиц и т. д.) по частным домохозяйствам. Наборы данных или файлы, базирующиеся на отдельных лицах как на единицах (объектах), используются для составления статистики (таблиц и т. д.) по лицам из частных домохозяйств.

113. Из вышесказанного очевидно, что существуют два основных типа файлов по результатам обследований домашних хозяйств: файлы домохозяйств и персональные файлы (по конкретным лицам). В большинстве случаев — это файлы домохозяйств в том смысле, что они содержат переменные показатели по домохозяйствам (переменные, связанные с наблюдениями единицы или объекта под названием «домохозяйство»). Некоторые из них представляют собой индивидуальные (персональные) файлы в том смысле, что они содержат значения переменных показателей по отдельным лицам (переменные, связанные с наблюдениями единицы или объекта под названием «человек»). Полные и окончательные файлы данных (наборы данных) содержат информацию по всем ответившим домохозяйствам и отдельным лицам из каждой обследованной первичной единицы выборки.

114. Приведенная ниже таблица 9.2 иллюстрирует, каким образом для дальнейшей обработки был реорганизован крупный файл Зимбабвийского межпереписного демографического обследования 1987 года. Во втором примере в таблице 9.3 представлены типовые файлы для обследования бюджета домохозяйств. Оба примера взяты из материалов работ Лагерлофа (1988 год) и Розена (1991 год).

Таблица 9.2
Файлы по домохозяйствам и отдельным лицам, использованные в Зимбабвийском межпереписном демографическом обследовании 1987 года

Файл	Тип файла	Содержание файла
Домохозяйство	Файл домохозяйства	Идентификатор домохозяйства (регион, провинция, район и т. д.) Ответы на все вопросы, связанные с домохозяйством Производные переменные величины, например размер домохозяйства (из файла членов домохозяйства) и т. д.
Человек	Персональный файл	Идентификатор домохозяйства (HID) плюс идентификатор личности (PID) Демографические характеристики: ВОЗРАСТ (AGE), ПОЛ (SEX), СЕМЕЙНОЕ ПОЛОЖЕНИЕ (MARSTAT), ПОСТОЯННЫЙ ЧЛЕН ДОМОХОЗЯЙСТВА (USMEM), РОДСТВЕННАЯ СВЯЗЬ С ГЛАВОЙ ДОМОХОЗЯЙСТВА (RELTH)
Умершее лицо	Персональный файл	Идентификатор домохозяйства (HID) плюс идентификатор умершего лица (DID) Данные умершего лица, которое являлось постоянным членом домохозяйства: ПОЛ (SEX), ВОЗРАСТ УМЕРШЕГО ЛИЦА (AGED)
Женщина в возрасте от 12 лет и старше	Персональный файл	HID, PID Данные для каждой женщины в составе домохозяйства в возрасте не младше 12 лет

115. Для целей составления таблиц двумерный файл необходим для большинства статистических программных пакетов. Многие из имеющихся программ общего назначения требуют данных в формате двумерного файла. В двумерном файле все записи имеют одинаковый набор переменных или полей и имеют одинаковую длину. Файл еще называется «плоским», поскольку для каждого респондента существует абсолютно одинаковый набор полей данных. Поля данных расположены одинаково в рамках каждой записи, и имеется фиксированное число записей идентичного внешнего вида. В таблице 9.4 показан двумерный файл, использованный для Зимбабвийского межпереписного демографического обследования 1987 года.

116. Файл домохозяйства содержит одну запись для каждого наблюдаемого домохозяйства, при этом каждая запись содержит следующую информацию:

- идентификатор домохозяйства;
- параметры плана выборки;
- наблюдаемые переменные показатели (домохозяйства);
- весовые переменные.

Идентификатор домохозяйства: комбинация *hrij* говорит о том, что домохозяйство *j* относится к *СУ i* в подразделении *r* страты *h*.

Таблица 9.3
Типовые файлы для обследования бюджета домохозяйств

Файл	Тип файла	Содержание файла
Домохозяйство	Файл домохозяйства	Идентификатор домохозяйства (регион, провинция, район и т. д.) Ответы на все вопросы, связанные с домохозяйством Производные переменные величины, например размер домохозяйства (из файла членов домохозяйства) и т. д.
Члены домохозяйства	Персональный файл	Идентификатор домохозяйства плюс идентификатор члена домохозяйства Демографические характеристики: возраст, пол, семейное положение, уровень образования и т. д. членов домохозяйства Информация об основных видах деятельности: статус занятости, род занятий и т. д.
Доход	Персональный файл	Идентификатор домохозяйства плюс идентификатор лица плюс идентификатор дохода Источник дохода:
Продукты питания	Файл домохозяйства	Идентификатор домохозяйства плюс идентификатор пищевого продукта Расходы на продукты питания:
Прочие товары краткосрочного пользования	Файл домохозяйства	Идентификатор домохозяйства плюс идентификатор товара Расходы на товары:
Товары долгосрочного пользования	Файл домохозяйства	Идентификатор домохозяйства плюс идентификатор товара длительного пользования Расходы на товары длительного пользования:
Сельское хозяйство	Файл домохозяйства	Идентификатор домохозяйства плюс идентификатор сельскохозяйственной позиции Расходы на сельское хозяйство:
Сельскохозяйственные средства производства	Файл домохозяйства	Идентификатор домохозяйства плюс идентификатор сельскохозяйственного средства производства Расходы на сельскохозяйственные средства производства:

Параметры плана выборки: в данном конкретном примере они таковы:

- S_h = 1982 год — число домохозяйств в страте выборки;
- a_h = размер выборки *СУ* в страте выборки;
- R_h = число подразделений, представленных в выборке из данной страты выборки;
- b_{hr} = число *СУ*, вошедших в выборку от каждого подразделения;
- S_{hi} =1982 год — число домохозяйств в данном *СУ*;
- M_{hi} =1987 год — число домохозяйств в данном *СУ*;
- m_{hi} = размер выборки домохозяйств из данного *СУ*.

Наблюдаемые переменные показатели: x, y, z обозначают переменные показатели домохозяйств.

Весовые переменные показатели: w обозначает переменный вес для данного домохозяйства.

Таблица 9.4
Формат двумерного файла, использованный в качестве файла домашнего хозяйства для Зимбабвийского межпереписного демографического обследования 1987 года

Идентификация			Параметры плана выборки								Переменные показатели			Вес переменной
Страта	Подразделение	СУ	H_j	S_h	a_h	R_h	b_h	S_h	M_h	m_{hi}	x	y	z	w
h	r	i	j								x_{hrij}	y_{hrij}	z_{hrij}	w_{hrij}

117. Организация персонального файла аналогична указанному выше файлу домохозяйства. Небольшим отличием является то, что идентификатором будет выступать идентификатор личности (PID), а индекс (k) будет относиться к отдельному лицу, в то время как «переменные показатели» будут означать переменные показатели по отдельным лицам.

118. Для целей распространения комплекты данных обследования необходимо систематизировать только по отдельным двумерным файлам (по одному файлу для каждого типа записи), поскольку формат для фиксированной длины собственной структуры также пригоден для перевода данных в стандартные системы управления базами данных (СУБД) для осуществления дальнейшей обработки данных или в стандартные статистические программы для составления таблиц и анализа. Перевод данных в системы управления базами данных не вызывает трудностей, так как собственная структура данных практически напрямую переводится в стандартный формат базы данных (ФБД), который воспринимается всеми ими в качестве вводных данных для отдельных таблиц [в этом случае идентификаторы записей служат в качестве естественных реляционных связок между таблицами (Munoz, 2003)].

9.4.3.5. Процедуры оценки и расчет весов

119. В главе 6 дано подробное описание логических обоснований и метода расчета весов для обследований домашних хозяйств (см. также ссылки на работы Розена в конце данной главы). Алгоритм расчета — от наблюдаемых значений до оценок статистических характеристик — называется процедурой точечной оценки. В качестве первого шага точечной оценки вес рассчитывается для каждого ответившего объекта. Затем рассчитываются «суммарные» значения путем суммирования взвешенных значений наблюдений (наблюдаемое значение, умноженное на соответствующий вес).

120. Муньос в своей работе (2003 год) дает надлежащее описание того, каким образом можно использовать компьютеризованную систему крупноформатных таблиц Microsoft Excel для осуществления процедуры формирования выборки при обследовании домашних хозяйств, предоставляя соответствующие инструкции на ее основных этапах: формирование инструментария первого этапа; отбор первичных единиц выборки с вероятностью, пропорциональной размеру; а также расчет вероятностей выборки и соответствующих весов выборки.

121. На практике построение взвешенных оценок — это довольно простой процесс. Начинать следует с набора данных первоначальной выборки и создавать новый набор данных путем умножения оценки каждого наблюдения на множитель, определяемый ее весом, затем надо использовать стандартные формулы для расчета того или иного параметра, используя набор взвешенных данных.

122. При этом следует отметить, что точные значения весов должны объединять в себе три элемента (Yansaneh, 2003), включая необходимые различные поправки (см. также главу 6). Базисные веса учитывают разброс вероятностей отбора между различными группами домохозяйств, как это предусмотрено в первоначальном плане обследования. Вторая поправка вносится на разброс долей неполучения ответов между областями или подгруппами обследования. И наконец, в некоторых случаях могут потребоваться постстратификационные поправки, для того чтобы данные обследования соответствовали показателям распределения, взятым из того или иного независимого источника, к примеру из последней переписи населения.

123. Другой фактор, усложняющий процесс расчетов, возникает в связи с повышенным спросом на статистические показатели на уровне областей обследования. Как рассматривается в главе 3, область обследования — это та подгруппа, по которой желательно получить отдельные оценки. Обычно они могут оговариваться на этапе составления плана выборки, однако их можно рассчитать и из производных данных. Кроме того, область обследования может быть одной стратой, комбинацией страт, административными регионами (провинциями, округами, городскими/сельскими районами и т. д.), а также может определяться с точки зрения демографических или социально-экономических характеристик (например, возраста, пола, этнической группы, уровня бедности и т. д.). Ниже делается попытка описать возможный способ построения наборов данных для облегчения оценки показателей по областям обследования.

124. Начнем с визуального ознакомления с файлом данных (наблюдений) (например, файла домохозяйств), такого как указанный выше файл для Зимбабвийского межпереписного демографического обследования. Этот файл содержит одну запись по каждому вошедшему в выборку домохозяйству. На завершающем этапе процесса обследования этот файл будет содержать следующую информацию по каждому домохозяйству:

a) идентификатор домохозяйства;

b) параметры выборки;

c) значения изучаемых переменных показателей x, y и z;

d) значение расчетного веса домохозяйства;

e) относится или нет данное домохозяйство к категории c;

f) относится или нет данное домохозяйство к области обследования g .

125. Данные информационные показатели (за исключением параметров выборки) обозначаются следующим образом:

- HID = идентификационный номер вошедших в выборку домохозяйств. Для простоты мы используем последовательную нумерацию 1, 2,…, n. Следовательно, n обозначает общий размер выборки.
- x, y и z — наблюдаемые значения переменных показателей X, Y, Z для данного домохозяйства.
- $c = 1$, если домохозяйство относится к категории c, в противном случае $c = 0$.
- $g = 1$, если домохозяйство относится к области обследования g, в противном случае $g = 0$.
- w = расчетный вес домохозяйства.

126. Значения индикаторных переменных показателей c и g обычно выводятся из значений других переменных и не наблюдаются напрямую. Например, мы можем определить категорию c, как

«живущие ниже черты бедности». Домохозяйствам не задается вопрос о том, принадлежат ли они к этой категории или нет. Такая классификация выводится, например, из данных по доходам домохозяйства и оговоренной черты бедности. Аналогичным образом, часто производные от других переменных требуются для определения того, принадлежит ли то или иное домохозяйство к конкретной изучаемой области *g* или нет (например, область обследования g может состоять из домохозяйств с тремя и более детьми). На этапе расчетов значения таких индикаторных показателей должны присутствовать в файле наблюдений.

127. Когда в файле наблюдений будут собраны все данные, он будет выглядеть так, как показано в таблице 9.5, ниже, за исключением того, что в нее не включаются параметры выборки.

128. Изложенные выше соображения были ограничены расчетом статистических характеристик для совокупности домохозяйств. Расчеты статистических характеристик для совокупности лиц осуществляются по тем же принципам. Как правило, расчетный вес для того или иного лица остается таким же, как и для домохозяйства, к которому данное лицо относится. Поскольку все члены типового домохозяйства перечислены в вопроснике, то или иное конкретное лицо включается в выборку лиц, если домохозяйство данного лица включено в выборку домохозяйств. Поэтому вероятность включения какого-либо лица в выборку такая же, как и вероятность включения домохозяйства, к которому это лицо относится. Следует отметить, однако, что вышеуказанное не соответствует действительности в случае формирования подвыборки среди домохозяйств. Например, в некоторых планах выборки процедура их формирования может предусматривать отбор только одного взрослого человека из одного домохозяйства или одного мужчину и одну женщину. В таких случаях вес отобранного(ых) лица (лиц) рассчитывается отдельно и не равняется весу домохозяйства.

129. Для полноты картины процедура расчета в обследованиях домашних хозяйств должна предусматривать предоставление оценок ошибок выборки (или стандартных ошибок) обследования, особенно для наиболее важных статистических показателей, которые специально определяются и публикуются. Глава 7 данного руководства целиком посвящена этому вопросу.

9.4.3.6. Составление таблиц, наборы данных для составления таблиц и баз данных

130. Существуют три основных базовых продукта статистического обследования (Sundgren, 1995):

- *Макроданные:* «статистические данные», дающие представление об определенных статистических характеристиках; получение таких данных является первоочередной целью проводимого обследования.
- *Микроданные: «наблюдения над отдельными объектами», лежащие в основе макроданных, получаемых в ходе обследования; эти данные необходимы для будущего использования и интерпретации результатов обследования.*
- *Метаданные: «данные, разъясняющие значение, точность, наличие и прочие важные характеристики соответствующих микро- и макроданных»; они необходимы для точного выявления* и извлечения профильных статистических данных по той или иной конкретной проблеме, а также для правильной интерпретации и (повторного) использования статистических данных.

Было бы целесообразным также рассмотреть схему составления многомерных (трехмерных) таблиц в целях обеспечения более гибкого доступа к результатам обследования, а также расширения возможностей ссылок на них, например, посредством веб-сайтов.

Таблица 9.5
Файл наблюдений с итоговыми данными для перемененных показателей обследования домашних хозяйств

HID	X	Y	Z	C	G	W
1	x_1	Y_1	z_1	c_1	g_1	w_1
2	x_2	Y_2	z_2	c_2	g_2	w_2
3	x_3	Y_3	z_3	cz_3	g_3	w_3
.	.	.		.	.	.
w.	.	.		.	.	.
N	x_n	y_n		c_n	g_n	w_n

131. Программа обследований домашних хозяйств должна в конечном счете обеспечить такое положение дел, когда архивы данных базируются на комбинации данных микро- и макроуровней. Для достижения этой цели необходимо иметь подробное описание структуры информации, собираемой с помощью многочисленных обследований.

132. Процедуру хранения данных следует рассматривать как трехэтапную операцию (Lundell, 2003):

- *введение в память*: в процессе ввода данные должны заноситься в память такими способами, которые в первую очередь оптимально работают с используемыми методами ввода и контроля данных, как отмечено выше;
- *хранение*: после того как данные введены и выверены, они должны добавляться в «хранилище», структура которого адаптирована к инструментам и методам анализа и распространения данных;
- *архивирование:* данные проекта должны архивироваться таким способом, который соответствует долгосрочным стандартам для обеспечения отсутствия в будущем каких-либо осложнений при извлечении данных.

133. Хранилище данных, содержащее достоверные данные, можно создать несколькими способами, используя один из перечисленных ниже методов (более подробную информацию по упоминаемым программным пакетам см. в добавлении к настоящей главе):

- двумерные файлы;
- реляционная база данных (например, «Сервер языка структурированных запросов (SQL) компании Microsoft»);
- статистические программы [например, «Система статистического анализа» (SAS) или «Статистический пакет для общественных наук» (SPSS)].

134. Для долгосрочного архивирования результирующих данных существует один основной вариант. Данные должны сохраняться в виде двумерных файлов в простом формате ASCII с приложенными описаниями записей. Большинство систем баз данных и статистических программ имеют функции экспорта данных в такие файлы без каких-либо проблем, а также могут легко импортировать данные из них.

9.5. Добавление

Варианты программного обеспечения для различных этапов обработки данных обследований

Тип операции	Варианты программного обеспечения
Система управления базой данных	Microsoft Structured Query Language (SQL) Server 2000, Standard Edition
	Microsoft Access
	Statistical Analysis System (SAS)
Ввод и редактирование данных	Visual Basic
	Microsoft Access
	Integrated Microcomputer Processing System (IMPS)
	Census and Survey Processing System (CSPro)
Извлечение данных	Statistical Analysis System (SAS)
	Statistical Package for the Social Sciences (SPSS)
	Microsoft Access
	Microsoft Excel
Составление таблиц, анализ и презентация данных	Microsoft Word
	Microsoft Excel
	Statistical Analysis System (SAS)
	Statistical Package for the Social Sciences (SPSS)
Расчет вариантности	CENVAR: компонент по расчету вариантности в составе системы ИСОДМ
	Computation and Listing of Useful STatistics on Errors of Sampling (CLUSTERS)
	Integrated System for Survey Analysis (ISSA)
	Survey Data Analysis (SUDAAN)
	Statistical Analysis System (SAS)
	Statistical Package for the Social Sciences (SPSS)
	Cluster Analysis and Regression Package (PC-CARP)

9.5.1. Microsoft Office

Программа Microsoft Office, которая была разработана корпорацией Microsoft Corporation представляет собой всеобъемлющий пакет, состоящий их различных компьютерных программ, в том числе:

- Microsoft Office Access — офисная программа управления базой данных, которая обеспечивает более простое использование и расширенные возможности импорта, экспорта и работы с файлами данных на языке XML (расширяемом языке разметки);
- Microsoft Office Excel — офисная программа работ с крупноформатными таблицами, которая поддерживает язык XML и предлагает новые функции, облегчающие анализ и обмен информацией;
- Microsoft Office Word — офисный текстовый редактор;
- Microsoft SQL Server 2000 — база данных сервера для полномасштабных деловых проектов и для обеспечения возможностей управления ресурсами;

- Microsoft Office Outlook — офисная программа персонального информационного администратора и коммуникации, предоставляющая единую платформу для управления электронной почтой, календарями и т. д.

Веб-сайт: www.microsoft.com/office/system/overview.mspx#EDAA

9.5.2. Visual Basic

Корпорация Microsoft выпустила программу Visual Basic в 1987 году. Visual Basic — это не только язык программирования, но и полновесная графическая среда проектирования. Эта среда позволяет пользователям с небольшим опытом программирования быстро разрабатывать прикладные программы Microsoft Windows, способные использовать технологию OLE (связывание и внедрение объектов) в отношении таких объектов, как крупноформатная таблица Excel. Кроме того, с помощью Visual Basic можно разрабатывать программы, которые могут использоваться как интерфейсная прикладная программа для системы базы данных, служащей в качестве пользовательского интерфейса для введения пользователем данных и выведения на дисплей отформатированных данных в более привлекательной и удобной форме, чем та, которую могут обеспечить многие версии программ в формате SQL.

Основным коммерческим аргументом Visual Basic является та легкость, с которой эта программа позволяет пользователю создавать привлекательные графические программы с минимальным программным кодированием. Основной объект в программе Visual Basic называется формой, что облегчает создание экранов для введения данных.

Веб-сайт: /www.engin.umd.umich.edu/CIS/course.des/cis400/vbasic/vbasic.html.

9.5.3. CENVAR

CENVAR — это компонент расчета вариантности в составе Интегрированной системы обработки данных для микрокомпьютеров (ИСОДМ), т. е. серия программных пакетов для ввода, редактирования, табулирования, расчета, анализа и распространения данных переписей и обследований. Система ИСОДМ разработана Бюро переписей Соединенных Штатов.

Веб-сайт: www.census.gov/ipc/www/imps/.

9.5.4. PC CARP

CENVAR базируется на программном обеспечении «Пакет программ кластерного анализа и регрессии для персональных компьютеров» (PC CARP), которое первоначально было разработано Университетом штата Айова. Для расчета вариантности в программе PC CARP используется процедура линеаризации.

Веб-сайт: www.census.gov/ipc/www/imps/.

9.5.5. Census and Survey Processing System (CSPro)

CSPro (Система обработки данных переписей населения и обследований) — это бесплатный общедоступный программный пакет для ввода, редактирования, табулирования и картографирования данных переписей и обследований. Программа CSPro была создана и внедрена совместными усилиями разработчиков ИСОДМ и ИССА: Бюро переписей Соединенных Штатов и компанией Serpro, S.A. Финансирование разработки этого пакета осуществлялось Отделом народонаселения

Агентства Соединенных Штатов по международному развитию. Программа CSPro разработана таким образом, чтобы со временем заменить и ИСОДМ, и ИССА.

Веб-сайт: www.census.gov/ipc/www/imps/

9.5.6. Computation and Listing of Useful Statistics on Errors on Sampling (CLUSTERS)

Пакет «Расчет и составление списков полезных статистических показателей по ошибкам выборки» (CLUSTERS) был первоначально разработан для расчета ошибок выборки для программы Всемирного обследования фертильности (ВОФ). Для расчета ошибок выборки в нем используется метод линеаризации по рядам Тейлора. Этот пакет также использовался для расчета ошибок выборки в различных обследованиях домашних хозяйств, особенно проводимых в рамках программ Обследований в области народонаселения и здравоохранения во многих развивающихся странах (см. работу Верма, 1982 год).

9.5.7. Integrated System for Survey Analysis (ISSA)

Пакет «Комплексной системы для анализа результатов обследования» (ИССА) разработан компанией Macro International Inc специально для программы Обследований в области народонаселения и здравоохранения. Этот пакет использовался во всех аспектах обработки данных, ввода данных, редактирования и табулирования. В нем также имеется модуль ошибок выборки, позволяющий производить расчет ошибок выборки для сложных демографических показателей, таких как уровни рождаемости и смертности, с применением метода «складного ножа» [см. публикацию компании Macro International Inc. (1996 год)].

9.5.8. Statistical Analysis System (SAS)

Система статистического анализа (SAS), разработанная компанией SAS, Inc. в 1966 году, является программным пакетом для анализа данных, управления файлами и расчета ошибок выборки [информацию по некоторым последним характеристикам пакета SAS см. в работе Энн и Уоттс (2001 год)].

9.5.9. Statistical Package for the Social Sciences (SPSS)

«Статистический пакет для общественных наук» (SPSS), разработанный компанией SPSS, Inc., является программным пакетом для анализа данных и обработки файлов [информацию по некоторым последним характеристикам этого пакета см. в публикации компании SPSS, Inc. (1988 год)].

9.5.10. Survey Data Analysis

Пакет «Анализ данных обследования» (SUDAAN), который был разработан научно-исследовательским институтом Triangle Institute (Research Triangle Part, штат Северная Каролина), представляет собой всеобъемлющий программный пакет для выборочных обследований (и связанных с ними данных) с мощными аналитическими функциями как для описательного анализа, так и аналитического моделирования. [Более подробную информацию можно получить из работы Шаха, Барнвелла и Билера (1996 год).]

Справочная литература и дополнительные источники информации

An, A., and D. Watts (2001). *New SAS Procedures for Analysis of Sample Survey Data.* SUGI paper, *No. 23*, Cary, North Carolina: SAS Institute, Inc.

Arnic, and others (2003). Metadata production systems within Europe: the case of the statistical system of Slovenia, Paper presented at the Metadata Production Workshop, Luxembourg. Eurostat Document 3331.

Australian Bureau of Statistics (2005). Labour statistics: concepts, sources and methods. Canberra: Statistical Concepts Library. Available from www.abs.gov.au/ Open Document Catalogue No. 6102.0.55.001.

Backlund, S. (1996). Future directions on IT issues. Mission report to National Statistical Centre, Lao People's Democratic Republic, Vientiane.

Brogan, D. (2003). Comparison of Data Analysis Software Suitable for Surveys in Developing Countries, United Nations Statistics Division, New York.

Central Statistics Office, Namibia (1996). The 1993/1994 National Household Income and Expenditure Survey (NHIES). Administrative and technical report, Windhoek: National Planning Commission.

Chromy, J., and S. Abeysasekara (2003). Analytical Uses of Survey Data, United Nations Statistics Division, New York.

Chronholm, P. and Edsfeldt (1996). Course and seminar on systems design, Mission report to Central Statistics (CSS), Pretoria.

Giles, M. (1996). *Turning Data into Information: A Manual for Social Analysis*. Canberra: Australian Bureau of Statistics.

Пол Глеввс (2005). «Обзор разработки вопросников для обследований домашних хозяйств в развивающихся странах» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования № 96. В продаже под № R.05.XVII.6.

Graubard, B., and E. Korn (2002). The Use of Sampling Weights in the Analysis of Survey Data, United Nations Statistics Division, New York.

International Labour Office (1990). *Survey of Economically Active Population, Employment, Unemployment and Underemployment: ILO Manual on Concepts and Methods*. Geneva: International Labour Office.

Jambwa, M. and L. Olsson (1987). Application of database technology in the African context. Invited paper, 46th session of International Statistical Institute, Tokyo.

Jambwa, M., C. Parirenyatwa, and B. Rosen (1989). Data processing at the Central Statistical Office: Lessons from recent history. Central Statistics Office, Harare.

Lagerlöf, Birgitta. (1988), Development of systems design for national household surveys. SCB R&D report, No. 4. Stockholm: Statistics Sweden.

Lehtonen, R., and E. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley & Sons.

Lundell, L. (1996). Information systems strategy for CSS. Report to Central Statistical Service (CSS), Pretoria.

_____ (2003). Census data processing experiences. Report to Central Bureau of Statistics (CBS), Windhoek.

Macro International, Inc. (1996). *Sampling Manual,* DHS-III Basic Document No. 6. Calverton, Maryland: Macro International, Inc.

Хуан Муньос (2003). «Руководство по обработке данных обследований домашних хозяйств» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96. В продаже под № R.05.XVII.6.

Olofsson, P. (1985). Proposals for Survey Design, Kingdom of Lesotho, Report on short-term mission on a labour-force survey to Bureau of Statistics, Maseru.

Olsson, Ulf. (1990)a. Approaches to agricultural statistics in developing countries: an appraisal of ICO's experiences, No. 12. Stockholm: SCB (Statistics Sweden, International Consulting Office). 20 July.

_____ (1990)b. Applied statistics lecture notes: special reports. TAN 1990:1. Stockholm Statistics Sweden International Consulting Office.

Ханс Петтерссон (2005). «Планирование инструментариев эталонной выборки и эталонных выборок для обследований домашних хозяйств в развивающихся странах» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96. В продаже под № R.05.XVII.6.

Puide, Annika, (1995). Report on a mission to Takwimu, Dar es Salaam, 21 November — 21 December 1994. TANSTAT 1994: 20 (20 January 1995). Stockholm: Statistics Sweden, International Consulting Office.

Rauch, L. (2001). Best Practices in Designing Websites for Dissemination of Statistics. Conference of European Statisticians Methodological Material. Geneva: United Nations Statistical Commission and Economic Commission for Europe.

Rosen, B. and B. Sundgren. (1991). Documentation for re-use of microdata from surveys carried out by Statistics Sweden, Working paper for Research and Development Unit, Statistics Sweden, Stockholm.

(2002)a. Mission on sampling: framework for the master sample, Kingdom of Lesotho. Report from a mission to the Bureau of Statistics, Maseru, Lesotho, 1–15 June 2002. LESSSTAT 2002:7. Stockholm: Statistics Sweden, International Consulting Office.

(2002)b. Report on the short-term mission on estimation procedure for master sample surveys. Maseru: Bureau of Statistics, Kingdom of Lesotho.

Rosen, Beugt. (1991). Estimation in the income, consumption and expenditure survey, ZIMSTAT 1991: 8:1.

Shah, B., B. Barnwell, and G. Bieler (1996). *SUDAAN User Manual*: *Release 7.0*, Research Triangle Park, North Carolina. Research Triangle Institute.

Педру Луиш ду Нашименту Силва (2005). «Составление отчетов о не обусловленных выборкой ошибках и их компенсация в обследованиях в Бразилии: текущая практика и будущие проблемы» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96. В продаже под № R.05.XVII.6.

SPSS, Inc. (1988), *SPSS/PC+V2.0 Base Manual*. Chicago, Illinois: SPSS.

St. Catherine, Edwin (2003). Review of data processing, analysis and dissemination for *Designing Household Survey Samples: Practical Guidelines*. United Nations Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Surveys, New York, 3–5 December 2003.

Sundgren, B. (1984). *Conceptual Design of Databases and Information Systems*, P/ADB Report E19. Stockholm: Statistics Sweden.

_____ (1986). *User-Oriented Systems Development at Statistics Sweden*. U/ADB Report E24, Stockholm: Statistics Sweden.

_____ (1991). *Information Systems Architecture for National and International Statistics Offices: Guidelines and Recommendations*. Geneva: United Nations Statistical Commission and Economic Commission for Europe.

_____ (1995). *Guidelines: Modelling Data and Metadata*. Geneva: United Nations Statistical Commission and Economic Commission for Europe.

Svensson, R. (1996). The Census Data Entry Application. Report from a mission to Central Statistical Service (CSS), Pretoria.

Thiel, Lisa Olson. (2001). Designing and developing a web site. Report from a mission to Bureau of Statistics, Maseru, Lesotho, 12–23 November 2001. LESSTAT: 2001:17. 28 December. Stockholm: SCB Statistics Sweden, Interntional Consultancy Office.

United Nations (1982). National Household Survey Capability Programme: survey data processing: a review of issues and procedures. DP/UN/INT-81-041/1. New York: United Nations Department of Technical Cooperation for Development Statistical Office.

_____ (1985). National Household Survey Capability Programme: household income expenditure surveys: a technical study. DP/UN/INT.88-X01/6E. New York: United Nations Department of Technical Cooperation for Development and Statistical Office.

Verma, Vijay. (1982). The estimation and presentation of sampling errors, World Fertility Survey, Technical Bulletins No. 11 (December). The Hague: International Statistical Institute. Voorburg, Netherlands.

Wallgren, Anders, and others (1996). *Graphing Statistics and Data: Creating Better Charts. Thousand Oaks, California:* Sage Publications, Inc.

World Bank (1991). The SDA survey instrument: an instrument to capture social dimensions of adjustment. Washington, D.C. Poverty and Social Policy Division, Technical Department, Africa Division.

Ибрагим С. Янсанех (2005). «Обзор вопросов планирования выборки для проведения обследований домашних хозяйств в развивающихся странах и странах с переходной экономикой» в публикации ДЭСВ ООН *Обследования домашних хозяйств в развивающихся странах и странах с переходной экономикой*. Методологические исследования, № 96. В продаже под № R.05.XVII.6.

Приложение I

Основы планирования выборки обследованиЯ

A.1. Введение

1. Выборка — это методология, с помощью которой отбирается часть совокупности, и результаты исследования такого сегмента обобщаются для всей той совокупности, часть или выборка из которой была отобрана. Как правило, применяются два вида выборки, а именно: вероятностная и детерминированная. Основной упор в этом руководстве сделан на вероятностных видах выборки. Данный обзор охватывает единицы обследования, план выборки, а также базовые стратегии формирования выборки с приведением примеров.

A.2. Единицы и концепции обследования

2. Начнем с определений единиц обследования и понятий, наиболее часто используемых в формировании выборки. *Элементы*: элементы (единицы) совокупности — это те единицы, по которым требуется получение информации. Они могут представлять собой элементарные единицы, составляющие ту совокупность, по которой необходимо сделать определенные выводы. Например, в обследовании рождаемости в домохозяйствах элементами конечного порядка обычно выступают женщины репродуктивного возраста. Для облегчения сбора данных в ходе обследования абсолютно необходимо, чтобы элементы были четко определены и легко поддавались выявлению.

3. *Совокупность:* совокупность — это суммарное количество определенных выше элементов. Таким образом, элементы представляют собой базисные единицы, составляющие и определяющие совокупность. Абсолютно необходимо определять совокупность с точки зрения:

- ее содержания, что требует определения типа и характеристик входящих в совокупность элементов;
- ее масштабов, которые означают географические границы, поскольку они связаны с охватом;
- времени, что означает период времени, в течение которого существует данная совокупность.

4. *Единицы наблюдения:* это те единицы, по которым собираются наблюдения. В обследованиях, проводимых с помощью опроса, они называются респондентами. *Отчетные единицы* — это элементы, предоставляющие требуемую информацию в ходе обследования. Отметим, что в некоторых случаях единицы наблюдения и отчетные единицы могут различаться. Например, в обследовании детей в возрасте до 5 лет информацию о своих детях обычно дают, в качестве посредников, родители от имени детей. В таких случаях включенные в выборку дети являются единицами наблюдения, в то время как родители — отчетными единицами.

5. *Единицы выборки:* единицы выборки используются в целях отбора элементов для включения в выборку. При поэлементной выборке каждая единица выборки содержит один элемент, в то время как, например, в гнездовой выборке единица выборки включает группу элементов, называемую кластером. Например, счетный участок (СУ) в качестве единицы выборки первого этапа будет содержать кластер домохозяйств. Иногда в одном и том же обследовании используются различные единицы выборки. Хорошим примером этого является многоэтапная выборка, в которой используется иерархическая структура единиц выборки (см. главу 3).

6. *Выборочные единицы:* отобранные единицы выборки могут называться выборочными единицами, а значения изучаемых характеристик таких выборочных единиц известны как выборочные наблюдения. *Единица анализа*: это единица, используемая на этапе составления таблиц и анализа. Такая единица может быть той или иной элементарной единицей или группой элементарных единиц. Следует отметить, что, как указывалось выше, единица анализа и отчетная единица не обязательно совпадают.

7. *Инструментарий выборки:* инструментарий выборки используется для выявления и включения в выборку единиц выборки, а также — в качестве основы для проведения оценок, базирующихся на данных выборки. Это предполагает, что та совокупность, из которой предполагается формирование выборки, должна быть представлена в физической форме. В идеальном случае инструментарий должен включать все относящиеся к определенной изучаемой совокупности единицы выборки с соответствующими реквизитами для идентификации. Инструментарии должны быть исчерпывающими и предпочтительно взаимоисключающими (более подробная информация приводится в главе 4). Широко используемыми типами инструментариев являются списочные, районные и множественные инструментарии.

8. *Списочный инструментарий: списочный инструментарий* содержит список единиц выборки, из которых может напрямую формироваться выборка. Предпочтительно, чтобы в данном инструментарии содержалась актуальная и точная информация по каждой единице выборки, такая как ее размер и другие характеристики. Дополнительная информация помогает при разработке и/или формировании эффективных выборок.

9. *Районный инструментарий:* районные инструментарии представляют собой многоэтапные инструментарии, широко используемые в обследованиях домашних хозяйств. В связи с этим данный инструментарий состоит из одного или нескольких этапов единиц выборки. Например, в двухступенчатом плане выборки инструментарий будет состоять из кластеров, которые могут называться первичными единицами выборки (ПЕВ); а в отобранных ПЕВ список домохозяйств становится инструментарием второго этапа. Как правило, инструментарии необходимы для каждого этапа отбора. Долговечность инструментария снижается по мере продвижения вниз по иерархической структуре единиц.

10. *Территориальные единицы:* территориальные единицы охватывают оговоренные сухопутные территории с четко определенными границами, в качестве которых могут выступать физические объекты, такие как дороги, улицы, реки, железнодорожные пути или воображаемые линии, служащие официальными границами между административными подразделениями. Счетные участки переписи обычно создаются в рамках самых мелких существующих в стране административных единиц. Это облегчает суммирование подсчитываемых показателей по административным единицам в области обследования.

11. *Инструментарий* или инструментарии, используемые в обследовании домашних хозяйств, должны давать возможность доступа ко всем единицам выборки обследуемой совокупности, с тем чтобы каждая единица выборки имела известную и отличающуюся от нуля вероятность

включения в выборку. Такой доступ может обеспечиваться путем осуществления выборки из инструментариев, обычно посредством двух или нескольких ступеней отбора. Инструментарий для первого этапа отбора должен включать все намеченные единицы выборки. На последующих этапах формирования выборки инструментарии необходимы только для тех единиц выборки, которые были отобраны на предыдущем этапе. Инструментарий выборки может храниться в виде бумажного документа и/или на электронном носителе.

A.3. План выборки

12. Как правило, под планом выборки подразумевается отбор и расчет выборки. Таким образом, вопрос состоит в том, как отобрать часть совокупности для включения в обследование. На практике план выборки предусматривает определение размера и структуры выборки, принимая во внимание расходы на проведение обследования. Наиболее предпочтительным является такой план выборки, который обеспечивает максимальную точность при оговоренном уровне расходов на обследование или минимальные расходы при оговоренном уровне точности.

13. Тем не менее, в самом начале необходимо подчеркнуть, что план выборки не может быть изолирован от других аспектов планирования и проведения обследования. В целом теория выборки касается того, каким образом для данной совокупности оценки при обследовании и ошибки выборки соотносятся с размером и структурой выборки.

A.3.1. Базовые требования к планированию вероятностной выборки

- Должна быть четко определена целевая совокупность;
- Должен иметься в наличии инструментарий выборки или инструментарии для многоэтапных выборок;
- Цели обследования должны быть однозначно оговорены с точки зрения содержательной части обследования, его аналитических переменных показателей и уровней разукрупнения (например, необходимы ли вам оценки или данные на национальном уровне, в разбивке по сельским/городским районам, на уровне провинций или уровне районов);
- Должны учитываться бюджетные ограничения и ограничения в части проведения работ на местах;
- Должны быть изложены требования к точности для определения размера выборки.

A.3.2. Значимость вероятностной выборки для крупномасштабных обследований домашних хозяйств

- Она позволяет охватить всю целевую совокупность посредством выборочного отбора;
- Она снижает погрешность выборки;
- Она позволяет обобщать результирующие показатели выборки применительно ко всей совокупности, из которой сформирована данная выборка;
- Она позволяет производить расчет ошибок выборки, являющихся показателем надежности данных;
- Было подтверждено, что она позволяет ведущей обследование организации представлять его результаты, не извиняясь за использование ненаучных методов.

A.3.3. Процедуры отбора, осуществления и оценки

- Каждый элемент совокупности должен быть представлен в инструментарии, из которого предполагается формирование выборки;
- Формирование выборки должно базироваться на случайном процессе, дающем каждой единице определенную вероятность отбора;
- Регистрироваться должны все и только все отобранные единицы;
- При оценке параметров совокупности на основе выборки данные по каждой единице (элементу) должны взвешиваться в соответствии с вероятностью отбора.

14. Случайный отбор единиц снижает шанс получения детерминированной выборки. Следовательно, рандомизация является надежным путем преодоления последствий непредвиденных искажающих факторов. Используемый метод создания выборки зависит от применяемой схемы формирования выборки. Более сложные планы выборки являются более жесткими к требуемым процедурам отбора.

A.4. Основы стратегий формирования вероятностной выборки

15. Существует целых ряд методологий вероятностной выборки, которые были разработаны в целях составления плана выборки, в том числе: простая случайная, систематическая, стратифицированная и гнездовая выборки. Ниже дается краткий обзор этих методов с приведением нескольких примеров.

A.4.1. Простая случайная выборка

16. Простая случайная выборка (ПСВ) — это метод формирования вероятностной выборки, в рамках которого каждый элемент совокупности имеет равный шанс/вероятность отбора. Выборка может формироваться с возможностью замены или без такой возможности. Этот метод редко используется в крупномасштабных обследованиях домашних хозяйств в силу высоких расходов на составление списков и передвижение. Он может рассматриваться как базовая форма вероятностной выборки, применимая в тех случаях, когда не имеется какой-либо предшествующей информации по структуре совокупности. ПСВ привлекательна за счет своей простоты с точки зрения процедур отбора и расчетов (например, ошибок выборки).

17. Несмотря на то что простая случайная выборка не очень широко используется, она лежит в основе теории выборки, прежде всего по причине простоты своих математических свойств. Таким образом, большинство статистических теорий и методологий исходит из простого случайного отбора элементов. И действительно, все остальные виды формирования вероятностной выборки могут рассматриваться как ограничения по использованию ПСВ, которые скрывают некоторые комбинации элементов совокупности. Простая случайная выборка служит двум целям:

- она устанавливает базовый уровень для сравнения относительной эффективности других методологий выборки;
- она может использоваться в качестве конечного метода отбора элементарных единиц в рамках более сложных планов, таких как планы многоэтапной гнездовой и стратифицированной выборок.

Приведенные ниже примеры иллюстрируют расчет вероятности отбора по принципу ПСВ:

1. В начале мы рассмотрим конечную совокупность из 100 домохозяйств $H_1, H_2, \ldots \ldots H_i, \ldots \ldots H_{100}$ с показателями доходов $X_1, X_2, \ldots \ldots X_i, \ldots \ldots X_{100}$.

В этом примере вероятность отбора любой конкретной единицы составляет $\frac{1}{100}$.

2. В качестве второго примера мы отметим, что для формирования выборки домохозяйств целевые домохозяйства могут нумероваться последовательно в инструментарии/списке. Используя случайные числа, может быть сформирована выборка, скажем, из 25 домохозяйств. При применении метода равновероятностной выборки (*Epsem*) f — это общая доля выборки для элементов.

Следовательно, $f = \frac{n}{N}$.

Если размер выборки $n = 25$, а общее число домохозяйств $N = 100$, то тогда доля выборки, т. е. вероятность отбора, составит
$\frac{25}{100} = \frac{1}{4}$.

A.4.1.1. Виды формирования выборки по принципу простой случайной выборки

18. Существуют два общепринятых метода формирования выборки по принципу простой случайной выборки, а именно:

a) простая случайная выборка с замещением (ПСВСЗ);

b) простая случайная выборка без замещения (ПСВБЗ).

Простая случайная выборка с замещением

19. Простая случайная выборка с замещением базируется на случайном отборе из совокупности, осуществляемом путем замещения выбранного элемента в совокупности после каждого составления выборки. Вероятность отбора того или иного элемента остается неизменной после каждого составления выборки, а любые сформированные независимые выборки остаются независимыми друг от друга. Эта особенность объясняет, почему ПСВ по умолчанию используется в качестве методологии формирования выборки во многих теоретических исследованиях в области статистики. Далее, поскольку допущение ПСВ значительно упрощает формулы для алгоритмов оценки, таких как алгоритмы оценки вариантности, этот вид выборки используется в качестве контрольного метода. В пункте 20, ниже, мы даем формулы расчета среднего значения (1) и вариантности (2) выборочного среднего при использовании простой случайной выборки с замещением. Эти формулы иллюстрируются числовыми примерами.

20. При выборке в n единиц, отобранных по методу выборки ПСВСЗ, для которой была собрана информация о переменном показателе x, среднее значение и вариантность рассчитываются по формулам:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i = \frac{1}{n}\left[x_1 + x_2 + \ldots \ldots + x_n\right]. \qquad \text{(A.1)}$$

1. Среднее значение

Когда x_1=24, x_2=30, x_3=27, x_4=36, x_5=31, x_6=38, x_7=23, x_8=40, x_9=25, x_{10}=32, тогда

$$\bar{x} = \frac{24 + 30 + 27 + \ldots \ldots + 25 + 32}{10} = 30{,}6.$$

2. Вариантность

$$V(\bar{x}) = \frac{s^2}{n}, \qquad \text{(A.2)}$$

где $s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_i^n x_i^2 - \frac{x^2}{n}\right] = \frac{1}{n-1}\left(\sum {x_i}^2 - n\bar{x}^2\right).$ **(A.3)**

$x^2 = (\sum x_i)^2 = 93636.$

Расчет при таких значениях дает

$$s^2 = \frac{(9684 - 9364)}{9} = 35{,}56$$

$$V(\bar{x}) = \frac{35{,}56}{10} = 3{,}56$$

$\mathrm{Se}(\bar{x}) = \sqrt{3{,}56}$.

Простая случайная выборка без замещения

21. Интуитивно лучше формировать выборку без замещения, поскольку в этом случае поступает больший объем информации, так как отсутствует вероятность повторения единиц выборки. Исходя из сказанного, стратегия простой случайной выборки без замещения — это наиболее часто применяемая процедура простой случайной выборки. В рамках данной процедуры процесс отбора продолжается до тех пор, пока не будут отобраны *n* отдельных единиц и не будут проигнорированы любые повторения. Это эквивалентно сохранению отобранной единицы (или единиц) и отбору любой дополнительной единицы с равной вероятностью из оставшихся единиц совокупности.

Ниже перечислены некоторые свойства простой случайной выборки без замещения:

- она дает фиксированный размер выборки;
- она дает в результате равную вероятность отбора каждого элемента/единицы (*Epsem*);
- как и при ПСВСЗ среднее значение и вариантность выборки являются неискаженными оценками параметров совокупности.

22. В пункте 24, ниже, мы даем формулы, используемые для расчета среднего значения и вариантности при использовании простой случайной выборки без замещения (4 и 5). Кроме того, мы даем числовые примеры в отношении способа вычисления среднего значения и вариантности выборки.

23. Допустим, что общее число начальных школ в регионе составляет 275. Из них формируется выборка 55 школ без замещения. Указанные ниже цифры представляют собой число работников (y_i) в каждой из отобранных школ.

5	10	32	6	8	2
15	16	35	7	50	6
2	6	47	20	20	6
7	6	35	6	16	2
21	2	48	4	15	2
7	5	46	6	7	
4	4	8	2	6	
7	2	7	8	2	
5	12	10	6	2	
2	40	7	7	19	

$\sum y_i = 688$, общее число работников;

$\sum y_i^2 = 18\,182$.

1. Выборочное среднее значение составляет

$$\bar{y} = \frac{\sum y_i}{n} \qquad \text{(A.4)}$$

где n — это размер выборки.

Расчет этих цифр дает

$$\bar{y} = \frac{688}{55} = 12{,}5.$$

2. Вариантность выборочного среднего значения составляет:

$$V(\bar{y}) = 1 - f \frac{s_y^2}{n}\ , \qquad \text{(A.5)}$$

где $1–f$ — это поправочный коэффициент совокупности и

$$s_y = 2\frac{1}{n-1}\left[\sum y_i \ -2n\bar{y}^2\right] = \frac{1}{54}[18\,182 - 8\,594\,] = 177{,}56\ , \qquad \text{(A.6)}$$

тогда

$$V(\bar{y}) = \left(1 - \frac{55}{275}\right)177{,}56/55 = 2{,}58 \text{ и } \mathrm{Se}\,(\bar{y}) = \sqrt{2{,}58}.$$

A.4.2. Систематическая выборка

24. Систематическая выборка — это метод формирования вероятностной выборки, при котором выборка формируется путем отбора каждого k-го элемента совокупности, где k представляет собой целое число больше единицы. Первый номер выборки должен отбираться в случайном порядке из первых k элементов. Отбор производится из порядкового списка. Это — распространенный метод отбора, особенно в тех случаях, когда имеется много единиц, и они последовательно пронумерованы от 1 до N. Предположим, что N (общее число единиц) является целым кратным требуемого размера выборки, n, и что k — это целое число, такое, чтобы $N = nk$. Затем выбирается случайное число между 1 и k. Предположим, что случайным начальным числом является 2, тогда выборка будет иметь размер n при следующей последовательной нумерации единиц:

$2, 2 + k, 2 + 2k, \ldots, 2 + (n - 1)K.$

Следует отметить, что выборка состоит из первой случайно выбранной единицы и каждой k-ой единицы вплоть до достижения требуемого размера выборки. Интервал k делит совокупность на кластеры или группы. В ходе этой процедуры мы отбираем один кластер единиц с вероятностью $1/k$. Поскольку первый номер отбирается в случайном порядке от 1 до k, каждая единица в предположительно равных по размеру кластерах имеет одинаковую вероятность отбора, $1/k$.

Рисунок A.1
Линейная систематическая выборка (*формирование выборки*)

1	2	**3**	4	5	6	7	**8**	9	10
11	12	**13**	14	15	16	17	**18**	19	20

A.4.2.1. Линейная систематическая выборка

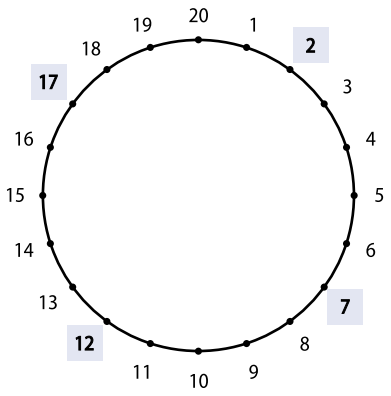
25. Если N, т. е. общее число единиц, является кратным числом желаемого размера выборки, иными словами, если $N = nk$, где n — желаемый размер выборки, а k — интервал выборки, тогда число единиц в каждой из возможных систематических выборок будет равно n. В такой ситуации процедура в рамках системы сводится к категоризации N единиц в k выборок по n единиц в каждой и к отбору одного кластера с вероятностью $1/k$. Когда $N = nk$, $\bar{y}$ — это неискаженная оценка среднего значения совокупности $\overline{Y}$. С другой стороны, когда N не является кратным размеру выборки, n, то число единиц, отобранных с применением метода систематической выборки с интервалом выборки, k, равным целому числу, ближайшему к N/n, не обязательно будет равняться n. Следовательно, когда N не равно nk, размеры выборок будут различаться, и выборочное среднее будет искаженной оценкой среднего значения совокупности. На рисунке A.1, выше, показано формирование выборки по методу линейной систематической выборки.

Приведенный выше пример иллюстрирует формирование выборки в 4 единицы из класса в 20 учащихся. Случайным стартовым номером является 3, N= 20, n = 4 и k=5. Получаемая в результате выборка состоит из единиц, имеющих номера 3, 8, 13 и 18.

A.4.2.2. Круговая систематическая выборка

26. Как отмечено выше, при линейной систематической выборке фактический размер выборки может отличаться от желаемого размера, при этом выборочное среднее будет искаженной оценкой среднего значения совокупности, когда N не является числом, кратным n. Однако метод круговой систематической выборки лишен вышеуказанного недостатка. По методу круговой систематической выборки позиции списка располагаются по окружности таким образом, чтобы за последней единицей списка следовала первая. Случайная стартовая позиция выбирается от 1 до N вместо выбора от 1 до k. Затем k-ая единица добавляется вплоть до отбора точно n элементов. При достижении конца списка отбор продолжается с его начала. На рисунке A.2 показывается формирование выборки по методу круговой систематической выборки, где N = 20, n = 4, k = 5, а случайная стартовая позиция равна 7. Таким образом, отобранными единицами стали номера 7, 12, 17 и 2.

Рисунок A.2
Формирование круговой систематической выборки



A.4.2.3. Проведение оценки в рамках систематической выборки

27. Указанные ниже формулы позволяют осуществлять расчет суммарного значения (7), выборочного значения (8) и вариантности (9), также даны числовые примеры, показывающие расчет оцениваемой совокупности выборочного среднего значения и вариантности.

1. Для оценки суммарного значения выборочное суммарное значение умножается на интервал выборки, следовательно

$$\hat{Y} = k\sum y_i . \qquad \text{(A.7)}$$

Оценка среднего значения совокупности равна:

$$\bar{y} = k\frac{\sum y_i}{N} . \qquad \text{(A.8)}$$

2. Расчет вариантности усложняется за счет того, что ее точную оценку невозможно вывести из единственной систематической выборки. Выход из этой ситуации заключается в допущении того, что нумерация единиц является случайной; в таком случае систематическая выборка может рассматриваться как случайная выборка. Следовательно, оценка вариантности для среднего значения рассчитывается по формуле:

$$V(\bar{y}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\sum s^2, \qquad \text{(A.9)}$$

где $s^2 = \frac{1}{n-1}\sum (y_i - y)^2$ и $y = \sum y_i$.

28. Точная оценка неискаженной вариантности систематической выборки может быть рассчитана путем формирования более чем одной систематической выборки из конкретной совокупности.

Числовые примеры

29. Предположим, что в какой-то провинции насчитывается 180 коммерческих ферм, имеющих по 30 и более голов крупного рогатого скота. Формируется выборка из 30 ферм с применением метода систематической выборки с интервалом $k = 6$.

Число голов крупного рогатого скота (y_i) в 30 отобранных фермах указывается ниже.

60	200	45	50	40	79	35	41	30	120
300	65	111	120	200	42	51	67	32	40
46	55	250	100	63	90	47	82	31	50

и $\sum y_i = 2\,542.$

1. Расчетное число голов крупного рогатого скота составляет
$\hat{Y} = k\sum y_i = 6 \times 2542 = 15252$.

2. Расчетное среднее число голов крупного рогатого скота на одну ферму составляет

$$\bar{y} = k\frac{\left(\sum y_i\right)}{N} = 6 \times 2542/180 = 84{,}7 \approx 85.$$

3. Вариантность выборочного среднего значения, рассчитываемого исходя из допущения случайной нумерации ферм, составляет::

$$V(\bar{y}) = 1 - f \frac{s_y^2}{n} \quad , \qquad \text{(A.10)}$$

где $s_y^2 = \frac{1}{n-1}\left\{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right\} = \frac{1}{29}(348\,700{,}00 - 215\,392{,}13) = 4596{,}80$,

таким образом, $V(\bar{y}) = (0{,}833)(153{,}227) = 127{,}64$

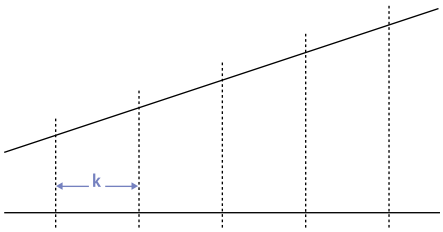
и $\text{Se}\left(\bar{y}\right) = \sqrt{127{,}64} = 11{,}30$.

30. Использование систематической выборки характеризуется рядом преимуществ и недостатков.

a) Преимущества:

- Отбор первой единицы определяет всю выборку. Это хорошо сочетается с работами на местах, поскольку крупные единицы выборки могут отбираться на местах регистраторами в процессе составления ими списков единиц.
- Выборка равномерно распределена по всей совокупности в случае надлежащей нумерации единиц в инструментарии. Тем не менее выборочная оценка будет более точной, если в совокупности наблюдается тот или иной тренд.
- Систематическая выборка обеспечивают неявную стратификацию. На рисунке A.3 ниже показана неявная стратификация посредством монотонного линейного тренда.

Рисунок A.3
Монотонный линейный тренд

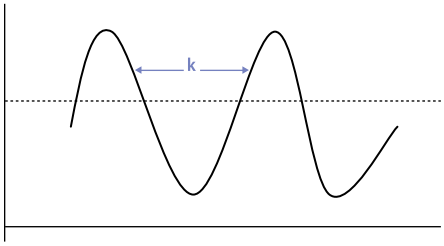


b) Недостатки:

- Если происходит периодическое изменение совокупности, систематическая выборка может дать заниженные или завышенные результаты. В таком случае интервал выборки попадает в линейную зависимость от данных. Например, если в течение 24 часов изучается транспортный поток на загруженной улице города и если выбранный интервал попадает на час пик, тогда будут постоянно получаться завышенные цифры. Вследствие этого исследование даст завышенные результаты. На рисунке A.4 показаны периодические изменения, которые могут привести к ненадежным оценкам при систематической выборке.
- Строго говоря, из единственной систематической выборки невозможно получить точную оценку вариантности.

- Такой метод отбора подвержен злоупотреблениям со стороны некоторых регистраторов/работников на местах.

Рисунок A.4
Периодические колебания



A.4.3. Стратифицированная выборка

31. Согласно методу стратифицированной выборки единицы выборки совокупности разделяются на группы, называемые стратами. Стратификация обычно осуществляется таким образом, чтобы совокупность подразделялась на разнородные группы, которые отличаются внутренней однородностью. Как правило, в случае когда единицы выборки однородны с точки зрения вспомогательного переменного показателя, называемого стратификационной переменной, обычно снижается вариантность оценки страты. Следует также отметить, что в стратификации наблюдается значительная гибкость в том смысле, что процедуры формирования выборки и расчетов могут отличаться между различными стратами.

32. Таким образом, при стратифицированной выборке группируются вместе единицы/элементы, которые более или менее аналогичны друг другу, так чтобы вариантность δ_h^2 в пределах каждой страты была небольшой. В то же время чрезвычайно важно, чтобы средние значения $(\overline{x}_h)$ различных страт максимально отличались друг от друга. Соответствующую оценку совокупности в целом можно получить, надлежащим образом комбинируя оценки изучаемой характеристики из различных страт.

A.4.3.1. Преимущества стратифицированной выборки

33. Основным преимуществом стратифицированной выборки является вероятное повышение точности оценок, а также возможность использования различных процедур формирования выборки в разных стратах. Кроме того, была установлена целесообразность применения стратификации в следующих ситуациях:

- при наличии совокупностей с несимметричным распределением во многих случаях при отборе из нескольких более крупных единиц могут использоваться более значительные доли выборки; это придает больший вес очень крупным единицам, и в конечном счете снижается вариантность выборки внутри страты;
- когда проводящая обследование организация имеет несколько местных отделений в различных регионах, на которые страна разделена в административных целях; в этом случае может оказаться целесообразным рассматривать эти регионы в качестве страт для облегчения организации работ на местах;
- когда требуются оценки, находящиеся в конкретных пределах погрешности не только для всего населения в целом, но и для определенных его подгрупп, таких как провинции,

городское или сельское население, население в разбивке по признаку пола и т. д.; такие оценки могут надлежащим образом обеспечиваться с помощью стратификации;

- если инструментарий выборки имеется только в виде подвыборок, которые могут предназначаться для регионов или оговоренных категорий единиц; в таком случае удобно и экономично с организационной точки зрения рассматривать такие подвыборки в качестве страт для формирования выборки.

A.4.3.2. Краткий обзор последовательности шагов при проведении стратифицированной выборки

- Вся совокупность единиц выборки разделяется на внутренне однородные, но внешне разнородные подсовокупности;
- внутри каждой страты формируется отдельная выборка из всех единиц выборки данной страты;
- для полученной по каждой страте выборки рассчитывается отдельное среднее значение по страте (или любой другой статистический показатель); затем, например, эти средние значения страт надлежащим образом взвешиваются для формирования комбинированной оценки среднего значения для совокупности;
- обычно пропорциональная выборка внутри страт используется тогда, когда целями обследования являются общие — например, национальные — оценки и когда обследование является многоцелевым;
- диспропорциональная выборка используется тогда, когда приоритет отдается областям обследования в виде подгрупп: например, в случаях, когда требуются одинаково надежные оценки для субнационального уровня.

A.4.3.3. Условные обозначения

34. Многие символы и их нижние индексы связаны со стратифицированной выборкой. Поэтому мы начинаем с определений некоторых общих обозначений и символов, используемых в рамках этой стратегии формирования выборки.

Показатели совокупности

Для H страт общее число элементов в каждой страте обозначается символами $N_1, N_2, \ldots \ldots N_h, N_H$.

Такая информация обычно неизвестна. Суммарное значение совокупности обозначается как

$$\sum_{h}^{H} N_h = N \, . \qquad \text{(A.11)}$$

Среднее значение страты обозначается как

$$\overline{X}_{hi} = \frac{1}{N} \sum_{i}^{N_h} X_{hi} = \frac{X_h}{N} \, , \qquad \text{(A.12)}$$

где символ X_{hi} — это значение h-го элемента в h-ой страте, а X_h — это сумма h-ой страты.

A.4.3.4. Веса

35. Веса обычно представляют собой пропорции элементов совокупности в страте и

$$W_h = \frac{N_h}{N}\ , \tag{A.13}$$

следовательно $\sum W_h = 1$,

$$S_h^2 = \frac{1}{N-1}\sum_{i=1}^{N_h}\left(X_{hi} - \overline{X}\right)^2. \tag{A.14}$$

A.4.3.5. Выборочные значения

36. Выборочное значение — это оценка, *рассчитанная из отобранных n_h элементов* в страте. В данном разделе мы даем описание общих символов, используемых для стратифицированной выборки:

a) для H страт допустим, что размеры выборки в каждой страте будут обозначаться как n_1, n_2, … … n_n, где

$\sum n_h = n$ — это общий размер выборки;

b) допустим, что x_{hi} — это элемент выборки i в страте h;

c) тогда

$$\overline{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$$ — это выборочное среднее для страты h; (A.15)

d) тогда

$$\overline{x}_{st} = \sum W_h \overline{x}_h$$ — это общее выборочное среднее; (A.16)

e) тогда

$$f_h = \frac{n_h}{N_h}$$ — это доля выборки для страты h; (A.17)

Вариантность n_h-го элемента в h-ой страте обозначается как

$$v(\overline{x}_h) = \sum\left[1 - \frac{n_h}{N_h}\right]\frac{s_h^2}{n_h}\ , \tag{A.18}$$

где s_h^2 — это вариантность элемента для h-ой страты, и она обозначается как:

$$s_h^2 = \frac{\sum \left(x_{hi} - \bar{x}_h\right)^2}{\left(n_h - 1\right)} \ . \qquad \text{(A.19)}$$

Вариантность выборочного среднего значения обозначается как :

$$v\left(\bar{x}_{st}\right) = \sum W_h^2 \left(1 - f_h\right) \frac{s_h^2}{n_h} \ . \qquad \text{(A.20)}$$

Ниже рассматриваются два вида стратегий стратифицированной выборки, а именно: пропорциональная и непропорциональная стратификации.

A.4.3.6. Пропорциональная стратификация

37. Пропорциональное распределение для стратифицированной выборки предусматривает использование единой доли выборки во всех стратах. Это означает, что одинаковая доля единиц отбирается из каждой страты. Например, если мы решим сформировать общую выборку в 10 процентов, это означает, что мы должны отобрать по 10 процентов единиц из каждой страты. Поскольку доли выборки во всех стратах одинаковы, вошедшие в выборку элементы будут различаться для разных страт. В пределах каждой страты размер выборки будет пропорционален числу элементов в страте.

В этом случае доля выборки определяется формулой $f_h = \frac{n_h}{N_h} = \frac{n}{N}$, что предполагает применение схемы равновероятностной выборки (*Epsem*).

Выборочное среднее значение обозначается как $\bar{x}_{st} = \sum W_h \bar{x}_h$. (A.21)

Вариантность выборочного среднего обозначается как $v\left(\bar{x}_{st}\right) = \frac{\left(1 - f\right)}{n} \sum W_h s_h^2$. (A.22)

A.4.3.7. Непропорциональная стратификация

38. Метод непропорциональной выборки предполагает использование различных долей выборки в разных стратах. Целью здесь является назначение таких долей выборки в страте, чтобы получить наименьшее значение вариантности общего среднего значения на удельную стоимость.

39. При применении этого метода доля выборки в каждой данной страте пропорциональна величине стандартного отклонения для этой страты. Это означает, что число единиц выборки, которое должно быть отобрано из любой страты, будет зависеть не только от общего числа элементов, но также и от стандартного отклонения вспомогательного переменного показателя.

При непропорциональном распределении вводится также обозначение функции стоимости. Например,

$$C = C_o + \sum c_h n_h \qquad \text{(A.23)}$$

где C_0 — это фиксированная стоимость, а c_h — это стоимость охвата выборки в какой-либо определенной страте.

Во многих ситуациях можно считать, что c_h является постоянной величиной для всех страт. Одной из широко используемых формул для непропорционального распределения выборок между стратами является распределение Неймана.

Там, где c_h является постоянной величиной, а $n = \sum n_h$, общий размер выборки будет фиксированным.

Число единиц, предполагаемых к отбору в рамках страт, обозначается как:

$$n_h = \frac{W_h s_h n}{\sum W_h s_h} \quad \text{or} \quad n_h = \frac{N_h s_h \cdot n}{\sum N_h s_h} \ . \tag{A.24}$$

Вариантность обозначается как:

$$v\left(\bar{x}_{st}\right) = \frac{\left(\sum W_h s_h\right)^2}{n} - \frac{1}{N}\sum W_h s_h^2. \tag{A.25}$$

Выражение, находящееся справа от знака минус, — это коэффициент поправки на конечность совокупности, который может быть отброшен, если выборка формируется из очень крупной совокупности, иными словами, если доля выборки невелика.

A.4.3.8. Общие замечания

- Значения совокупности, S_h и C_h, как правило, неизвестны; вследствие этого оценки могут выводиться из предшествующих или экспериментальных выборочных обследований;
- непропорциональное распределение не очень эффективно для отбора пропорций;
- в случае многоцелевых обследований могут возникать конфликты в отношении переменных величин, подлежащих оптимизации;
- как правило, непропорциональное распределение в результате дает наименьший уровень вариантности.

40. Приводимые ниже примеры иллюстрируют расчет размеров и вариантности выборки по методам пропорциональной и непропорциональной стратификации. В этом гипотетическом примере школы стратифицированы на основе числа наемных работников. Общее число начальных школ в какой-то провинции составляет 275. Формируется выборка из 55 школ, которая затем стратифицируется исходя из числа наемных работников.

A.4.3.9. Определение размеров выборки в рамках страты

41. Следует обратиться к таблице 1, ниже.

Пропорциональное распределение

Для пропорционального распределения используется общая доля выборки.

Следовательно, $\frac{n}{N} = f$ — это доля выборки в генеральной совокупности, применяемая к общему числу единиц в страте.

В приведенном выше примере $f = \frac{55}{275} = 0{,}2$, или 20 процентов.

Распределение размеров выборки дается в столбце 4 таблицы A.1; например, размер выборки для страты 1 составляет $n_h = 0{,}2 \times 80 = 16$.

Непропорциональное распределение

Для получения размеров выборок по различным стратам используется следующая формула:

$$n_h = \frac{W_h s_h}{\sum W_h s_h}(n). \qquad \text{(A.26)}$$

Например, $n_h = \frac{0{,}3750}{2{,}4474} \times 55 = 8$ для страты 1.

Остальные результаты приводятся в столбце 5 таблицы.

Таблица A.1
Число школ в разбивке по числу наемных работников

Страта	Число работников на каждую отобранную школу (y_{hi})	Общее число школ в каждой страте (N_h)	Число отобранных школ в разбивке по стратам		W_h	s_h^2	s_h	$W_h s_h$	$W_h s_h^2$
			Пропорциональное распределение (n_h)	Непропорциональное распределение (n_h)					
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	2,4,2,2,4, 2,2,4,2,2, 2,2,2,2,5,5	80	16	8	0,2909	1,663	1,289	0,3750	0,48
2	7,7,7,6,8, 7,7,6,7,6 6,8,6,7,8, 6,7,6,6,6	100	20	6	0,3636	0,537	0,733	0,2665	0,19
3	10,12,10,15, 21,16,20,20, 16,19,15	55	11	18	0,2000	15,564	3,945	0,7890	3,11
4	32,35,35,48, 46,47,50,40	40	8	23	0,1455	48,836	6,989	1,0169	7,10
Всего		**275**	**55**	**55**	**1,0000**			**2,4474**	**10,90**

Примечание: N = общее число начальных школ
n = общее число начальных школ во всей выборке
N_h = размер h-ой страты
n_h = размер выборки h-ой страты.

A.4.3.10. Расчет вариантности

42. Расчет вариантности при пропорциональном и непропорциональном распределении иллюстрируется применением формул A.27 и A.28, соответственно.

Пропорциональная стратификация:

$$V\left(\bar{y}_{prop}\right)=\frac{1-f}{n}\sum w_h s_h^2=\frac{(1-0{,}2)}{55}(10{,}9)=0{,}16\,. \tag{A.27}$$

Непропорциональная стратификация:

$$V\left(\bar{y}_{opt}\right)=\frac{\left(\sum w_h s_h\right)^2}{n}-\frac{1}{N}\sum w_h s_h^2=\frac{(2{,}4474)^2}{55}-\frac{10{,}9}{275}=0{,}07. \tag{A.28}$$

A.4.3.11. В общем случае

$$v\left(\bar{x}_{st}\right)OP\le v\left(\bar{x}_{st}\right)PROP\le v\left(\bar{x}_{st}\right)\le\left(\bar{x}_{st}\right)SRS. \tag{A.29}$$

A.4.4. Гнездовая выборка

43. В предыдущих разделах обсуждения касались тех методов выборки, согласно которым предполагалось, что элементарные единицы выборки располагаются в соответствии со списком инструментария; и схема была такова, что отдельные единицы могли отбираться непосредственно из инструментария. При гнездовой выборке единицы отбора более высокого уровня, например счетные участки (см. главу 3), содержат несколько элементарных единиц. В этом случае единицей выборки является кластер. Например, простой метод отбора случайной выборки домохозяйств в городе может предусматривать наличие списка домохозяйств. Это может оказаться невозможным, поскольку на практике может и не существовать полного инструментария по всем домохозяйствам в данном городе. Для того чтобы обойти эту проблему, могут формироваться кластеры в виде городских кварталов. Затем может быть сформирована выборка из этих кварталов, а после этого может быть создан список домохозяйств в вошедших в выборку кварталах. При необходимости из каждого квартала может быть сделана выборка домохозяйств в размере, например, 10 процентов.

A.4.4.1. Причины использования гнездовой выборки

44. Ниже приводятся некоторые соображения в пользу применения гнездовой выборки, особенно в многоэтапных планах выборки.

- Использование кластеров сокращает командировочные и прочие расходы, связанные со сбором данных.
- Использование кластеров может улучшить надзор, контроль, последующий охват и другие аспекты, влияющие на качество собираемых данных.
- Создание инструментария оказывается менее затратным, так как проводится поэтапно. Например, при многоэтапной выборке, как рассматривается в главе 3, инструментарий,

охватывающий всю совокупность, требуется только для отбора на первом этапе ПЕВ, или, иными словами, кластеров. На любом более низком этапе инструментарий потребуется только в пределах единиц, отобранных на предыдущем этапе.

- Кроме того, инструментарии для более крупных единиц на более высоком этапе, как правило, более долговечны и, вследствие этого, применимы в течение более длительного периода времени. Списки малых единиц, таких как домохозяйства и, в частности, отдельные лица, имеют тенденцию к устареванию за короткие сроки.
- Существуют преимущества административного характера при проведении обследования.

45. В общем, следует отметить, что при сравнении гнездовой выборки с элементной выборкой того же размера мы обнаружили, что в гнездовой выборке уровень затрат на элемент более низкий из-за более низкой стоимости составления списков и/или размещения элементов. С другой стороны, вариантность элементов более высокая вследствие неодинаковой однородности элементов (внутриклассовая корреляция) в кластерах. Мы иллюстрируем базовую гнездовую выборку посредством рассмотрения одноэтапного плана выборки (многоэтапные планы выборки были представлены и подробно обсуждались в главе 3).

A.4.4.2. Одноэтапная гнездовая выборка

46. В каком-то конкретном районе может оказаться практически неосуществимым получить список всех домохозяйств, а затем сформировать из них выборку. При этом, однако, может существовать возможность найти список деревень, подготовленный в ходе предыдущего обследования или сохраняемый в административных целях. В этом случае мы получим выборку деревень, затем получим информацию обо всех домохозяйствах в отобранных деревнях. Такая схема представляет собой план одноэтапной гнездовой выборки, поскольку после формирования выборки деревень охватываются все единицы кластера — в данном случае домохозяйства.

47. Формирование выборки с помощью кластеров можно проиллюстрировать следующим образом. Предположим, что из совокупности деревень (кластеров) формируется выборка с равной вероятностью отбора. При одноэтапной гнездовой выборке в эту выборку будут включаться все деревни из отобранных деревень.

При условии, что

A = общее число деревень

B = общее число домохозяйств в кластере

a = выборка деревень

и поэтому,

$aB = n$ представляет собой число элементарных единиц (домохозяйств) в общей выборке

и

$AB = N$ — это общее число домохозяйств во всех деревнях,

тогда вероятность отбора элемента с равной вероятностью рассчитывается по формуле

$$\frac{a}{A}\times\frac{B}{B}=\frac{n}{N}=f, \qquad \text{(A.30)}$$

где N — это общее число элементарных единиц, а f — доля выборки. В этом случае вероятность отбора равна просто $\frac{a}{A}$.

A.4.4.3. Формулы расчета среднего значения и вариантности выборки

48. Ниже приводятся формулы расчета выборочного среднего значения и вариантности выборки:

Выборочное среднее значение:

$$\bar{y} = \frac{1}{aB} = \sum_{\alpha=1}^{\alpha} \sum_{\beta=1}^{\beta} \bar{y}_{\alpha\beta} = \frac{1}{a} \sum_{\alpha=1}^{\alpha} \bar{y}_{\alpha} . \quad \textbf{(A.31)}$$

Выборочное среднее является неискаженной оценкой среднего значения для совокупности:

$$E(\bar{y}) = \frac{1}{A} \sum_{\alpha=1}^{a} \bar{y}_{\alpha} = \bar{Y}. \quad \textbf{(A.32)}$$

Фактически, поскольку размер выборки является фиксированным ($aB = n$), а отбор проводится с равной вероятностью, тогда среднее значение ($\bar{y}$) — это неискаженная оценка среднего значения для совокупности $\bar{Y}$.

Вариантность

Если кластеры формируются путем простого случайного отбора, вариантность можно рассчитать следующим образом:

$$V(\bar{y}) = (1 - f)s_{\alpha}^2 , \quad \textbf{(A.33)}$$

$$\text{где } s_{\alpha}^2 = \frac{1}{a-1} \sum_{\alpha=1}^{a} (\bar{y}_{\alpha} - \bar{y})^2.$$

49. Важно отметить, что эти значения свободны от ошибки выборки, поскольку они базируются на значениях всех элементов в B, а не в выборке. Вариантность выборочного среднего значения обусловлена лишь различиями между средними значениями для кластеров.

Приложение II

Список экспертов

Список экспертов, принимавших участие в совещании группы экспертов Организации Объединенных Наций по рассмотрению проекта Руководства по планированию выборочных обследований домашних хозяйств (Нью-Йорк, 3–5 декабря 2005 года)[a]

Фамилия	Должность и место работы
Оладеджо Ойелеке Аджайи	Консультант по вопросам статистики, Нигерия
Беверли Карлсон	Отдел по вопросам производства, производительности и управления, Экономическая комиссия для Латинской Америки и Карибского бассейна, Сантьяго, Чили
Самир Фарид	Консультант по вопросам статистики, Египет
Мафион М. Джамбва	Технический консультант, Сообщество по вопросам развития стран юга Африки/Европейский союз, Габороне, Ботсвана
Удая Шанкар Мишра	Член-корреспондент, Гарвардский университет, Бостон, Массачусетс, Соединенные Штаты Америки
Ян Кордос	Профессор, Варшавская школа экономики, Варшава, Польша
Эдвин Ст. Катерин	Директор, Национальное статистическое управление, Сент-Люсия
Энтони Тёрнер	Консультант по вопросам выборки, Соединенные Штаты Америки
Шиам Упадхайя	Директор, Объединенная статистическая служба (ИНСТАТ), Непал
Ибрагим Янсанех	Заместитель начальника, Отдел по вопросам стоимости жизни, Комиссия по международной гражданской службе, Организация Объединенных Наций, Нью-Йорк, Соединенные Штаты Америки

a Доклад совещания группы экспертов см. в документе ESA/STAT/AC.93/L.4.