ОБЕСПЕЧЕНИЕ КАЧЕСТВА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ

О РАБОТАХ ПО ВЫЯВЛЕНИЮ В СТАТИСТИЧЕСКОМ РЕГИСТРЕ ХОЗЯЙСТВУЮЩИХ СУБЪЕКТОВ «ЛОЖНО» АКТИВНЫХ ЕДИНИЦ

Р.П. Айчепшева,
Территориальный орган Росстата по Карачаево-Черкесской Республике

Одними из важнейших составляющих в динамичном процессе развития и совершенствования современной статистической системы являются повышение качества и достоверности представляемой пользователям статистической информации, а также снижение информационной нагрузки на хозяйствующие субъекты, предполагающее преимущественно выборочные методы обследования респондентов, переход на которые в соответствии с Федеральной целевой программой «Развитие государственной статистики России в 2007-2011 годах» отнесен к приоритетному направлению деятельности.

В этих условиях задача формирования надежной официальной статистической информации о социальном, экономическом, демографическом и экологическом положении регионов Российской Федерации, а также страны в целом становится довольно сложной для исполнения и требует дополнительного оснащения органов государственной статистики такими информационными ресурсами, как данные из административных источников, а также показатели, которые можно сформировать по результатам периодических экономических переписей, проводимых сплошным методом.

Необходимо отметить, что в целях обеспечения возможности получения и применения в статистической практике каждого из указанных дополнительных источников ведется планомерная работа как на государственном уровне, так и на уровне Росстата и заинтересованных ведомств, состоящая прежде всего из подготовки и принятия соответствующих нормативно-правовых актов, обеспечивающих реализацию в ближайшей перспективе различных актуальных задач, в том числе и такой, как проведение всеобщей экономической переписи субъектов малого предпринимательства. Общеизвестно, что затягивание упомянутого мероприятия имеет негативные последствия при формировании эффективной политики государства в сфере поддержки малого предпринимательства, а также влечет за собой такую проблему, как отсутствие достаточной информации, позволяющей оценить реальные параметры данного сектора, значимость которого для развития экономики регионов и России в целом не вызывает никакого сомнения.

В условиях же реализации Федерального закона № 209-ФЗ от 24.07.2007 «О развитии малого и среднего предпринимательства в РФ» решение вопроса о проведении экономической переписи приобретает особенное значение. Связано это с тем, что обоснованное и адекватное направление помощи субъектам малого бизнеса должно основываться на достоверных данных о масштабах его деятельности и тенденциях развития.

Кроме того, применение вышеупомянутого выборочного метода наблюдения должно базироваться на уточненной списочной основе, а получение последней возможно, опять же, только по результатам проведения периодических сплошных переписей хозяйствующих субъектов, так как формирование качественных итогов при подобных обследованиях напрямую зависит от состояния исходных данных той совокупности объектов, на основе которой формируется выборка.

В этой связи особое значение приобретает принятие Правительством РФ 14.02.2009 г. распоряжения № 201-р, регламентирующего проведение во II квартале 2011 г. сплошного федерального статистического наблюдения за деятельностью субъектов малого и среднего предпринимательства.

В рамках подготовки к осуществлению данного крупномасштабного мероприятия, а также в целях повышения качества Статистического регистра хозяйствующих субъектов (Статрегистра), служащего в соответствии с действующей методологией информационной основой для проведения различных обследований и наблюдений, Карачаево-Черкесскстат в течение последнего времени проводит целенаправленную работу по мониторингу объектов, содержащихся в административном и статистическом разделах Статрегистра по следующим направлениям:

- полнота учета хозяйствующих субъектов, а также уточнение структуры юридических лиц;

- достоверность идентификации объектов учета классификационными признаками, а также установление типологии по объектам, ранее не представлявшим статистическую отчетность;

- актуальность информации об адресах местонахождения и контактных линиях;

- полнота охвата объектов статистической и финансовой отчетностью;

- выявление «ложно» активных единиц.

Разработка каждого из перечисленных направлений состоит в свою очередь из отдельных блоков, результатом выполнения которых станет формирование в Карачаево-Черкесскстате информационного ресурса, включающего сведения в разрезе каждого предприятия о его деловой активности, фактическом адресе и контактных линиях.

Данный ресурс предусматривается использовать в качестве вспомогательного источника по хозяйствующим субъектам, которые будут участвовать в сплошной экономической переписи 2011 г., а также для установления четких параметров той части Статрегистра, которая относится к категории «спящих» объектов.

В целях более предметного рассмотрения некоторых из упомянутых блоков обратимся к информации о работах, осуществляемых в Карачаево-Черкесскстате помимо стандартных процедур, регламентированных инструктивными материалами по ведению Статрегистра, в том числе:

1. Обеспечение полноты и актуальности учета хозяйствующих субъектов, включающее такие мероприятия, как проведение: ежегодных работ по сверке с ЕГРЮЛ; сверок (*в соответствии с разработанным планом*) с различными ведомствами по уточнению перечней предприятий, входящих в их структуру; единовременной сверки филиалов, представительств и структурных подразделений юридических лиц (*расположенных на одной и разных территориях с головными организациями*) с перечнем данных объектов, состоящих на налоговом учете; периодических сверок с головными организациями по уточнению сведений по их местным единицам;

2. Обеспечение достоверности идентификации объектов классификационными признаками, а также установление типологии по объектам, ранее не представлявшим статистическую отчетность, на основе работ по: анкетированию вновь созданных и прошедших перерегистрацию объектов в процессе доведения до них Уведомлений о кодах ОК ТЭИ, а также при их непосредственных обращениях в Карачаево-Черкесскстат; взаимодействию с акционерными обществами республики по выяснению актуальных сведений об изменениях в составе их участников и распределении долей в уставном фонде;

3. Обеспечение актуальности информации о фактических адресах расположения предприятий и их контактных линиях, которое также основывается на анкетировании вновь созданных и прошедших перерегистрацию предприятий и организаций; кроме того, используются сведения, полученные по результатам проводящихся рейдов по обходу предприятий, осуществляемых в основном специалистами районных подразделений Карачаево-Черкесскстата; внесение данных о дислокации мест деятельности с применением административных источников, включая информацию, предоставляемую лицензирующими органами; осуществление работ по взаимодействию со справочной телефонной службой по уточнению номеров телефонов и по внесению сведений об электронных адресах по объектам, предоставившим упомянутую информацию в ходе организации работ по переходу предприятий на электронный способ представления отчетности.

Перечисленные выше мероприятия по повышению качества Статрегистра позволяют осуществить: уточнение классификационных признаков и типов вновь созданных предприятий в генеральной совокупности (ГС); актуализацию данных по обособленным подразделениям; накопление сведений о фактических местах деятельности или местах нахождения хозяйствующих субъектов; обновление информации о контактных линиях; выявление в Статрегистре «ложно» активных единиц (*работа, тесно взаимосвязанная с другим направлением - обеспечением полноты охвата объектов генеральной совокупности статистической и финансовой отчетностью*), к которым относятся объекты Статрегистра, числящиеся в нем экономически активными, а фактически являющиеся бездействующими.

Работы по выявлению «ложно» активных единиц ведутся в Карачаево-Черкесскстате на постоянной основе, в том числе при формировании статистического раздела регистра на очередной отчетный год, когда структурными подразделениями Карачаево-Черкесскстата подвергаются анализу объекты, не представляющие длительное время отчетность, и формируются перечни объектов, подлежащих удалению из генеральной совокупности. Данная процедура осуществляется также и при получении справок, подтверждающих информацию об отсутствии деятельности у предприятий по итогам за предыдущий год, или перечней объектов, не обнаруженных по адресу.

Удаление из Статрегистра «ложно» активных единиц осуществляется и по результатам сверок информационных массивов Статрегистра с ЕГРЮЛ в полном объеме, когда обнаруживаются юридически ликвидированные объекты, информация о прекращении деятельности которых своевременно не представлялась в Карачаево-Черкесскстат регистрирующими органами республики.

Кроме того, начиная с 2008 г. и по настоящее время городскими и районными подразделениями Карачаево-Черкесскстата на периодической основе проводятся обходы тех предприятий, которые при предварительном рассмотрении можно отнести к «ложно» активным единицам. Данная работа состоит из нескольких этапов.

Учитывая, что выявление «ложно» активных единиц имеет наибольшее значение при работе со статистическим разделом регистра, на первом этапе была

подвергнута обследованию совокупность объектов, относящихся к кругу коммерческих организаций, из числа тех, которые длительное время не представляют статистическую и финансовую отчетность и содержатся в генеральной совокупности под соответствующими метками («22», «12») в поле 23 (*признак наличия годовой бухгалтерской и статистической отчетности*). Количество таких организаций составило 838 единиц, что соответствовало 25% от общего числа коммерческих предприятий ГС-08 (3367 единиц).

В целях оптимизации намеченной работы данные объекты были сопоставлены с ранее созданными в Карачаево-Черкесскстате базами данных: «Деятельность не ведется» (*формируется по объектам, предоставившим справки об отсутствии деятельности в предыдущем году или нулевую отчетность за аналогичный период*) и «Не найдены по адресу» (*формируется в рамках работ по доведению Уведомлений, а также при работе по охвату объектов финансовой и статистической отчетностью*), а также сопоставлены с административной частью регистра с целью исключения из списка удаленных объектов и предприятий, находящихся в стадии ликвидации.

В результате перечень объектов уменьшился на 318 объектов и составил 520 единиц (15% от числа коммерческих предприятий ГС). Этот перечень и подлежал изучению методом ареолярного контрольного обследования силами районных подразделений Карачаево-Черкесскстата.

По итогам проведенной работы, были получены следующие данные: 353 предприятия (68% от числа обследуемых) были обнаружены по адресу, из которых 222 объекта (43%) в период проведения обследования охвачены годовой статистической или финансовой отчетностью (*это в основном вновь созданные в 2006-2007 гг. организации*).

Остальные 131 единица (25%) из перечня найденных по адресу предприятий или предоставили справки о том, что деятельность приостановлена [100 единиц (19%)] (*эти объекты на данном этапе не исключены из ГС, но взяты под контроль в целях рассмотрения вопроса об их удалении при формировании ГС на 2010 г.*), или в одном случае отказывались идти на взаимодействие со специалистами, проводящими обследование, а в другом - вводили их в заблуждение, обещав отчитаться, но в последующем не предоставили о себе никакой информации [31единица (6%)].

К числу не обнаруженных по адресу отнесено 83 объекта (16%), из которых 63 (12%) по согласованию с отраслевыми отделами удалены из ГС-2009, а остальные 20 предприятий взяты под контроль (*по аналогии с предыдущим пунктом*), так как являются относительно новыми объектами.

По оставшимся 87 объектам (17%) сложилась такая картина - с 44 (8%) из них нет возможности связаться или отнести объекты к категории не обнаруженных по адресу (*наличие кодовых замков при входе или заброшенных объектов, постоянное отсутствие жильцов по указанному адресу и т. д. - данные объекты также взяты под контроль*), а 35 (7%) - остались на данный момент необследованными.

Следующим этапом было предусмотрено обследование коммерческих предприятий, отсутствующих в ГС, как длительное время не представляющих отчетность, но продолжающих состоять на учете в административном разделе регистра. Число отобранных для изучения объектов составило 561 единицу, из которой 174 (31%) - не были обнаружены по адресу регистрации, а 23 (4%) - постоянно отсутствовали по официальному месту нахождения предприятия. По семи (1%) предприятиям выяснено, что их руководители умерли, но предприятия продолжают содержаться в ЕГРЮЛ в статусе действующих объектов (*данная информация, наряду с перечнями организаций, не обнаруженных по адресу, была направлена для сведения и использования при ведении ЕГРЮЛ в адрес регистрирующих органов республики*).

Из 171 (31%) объекта, обнаруженного по адресу, по данным на 01.07.2009, 95 (17%) заявили, что деятельности не ведут и не планируют ее вести, а 76 (14%), напротив, информировали обследовавших их специалистов о том, что деятельность или постоянно ведется, или планируется в ближайшее время (*из них 43 объекта были включены в ГС-2009*). Оставшиеся 186 (33%) объектов находятся в работе.

В дальнейшем планируется осуществление аналогичного обследования также и по кругу объектов, относящихся к некоммерческим организациям.

При этом следует отметить, что промежуточные результаты проводимых в Карачаево-Черкесскстате работ (*часть объектов на текущий момент времени все еще остается необследованной*) по микропереписи коммерческих предприятий, содержащихся в ГС с метками «22» и «12», а также объектов, отсутствующих в статистическом разделе, позволяют сформировать некоторые предварительные выводы об уровне качества территориального раздела регистра с учетом «Рекомендаций по регистрам», разработанных Рабочей группой, включающей специалистов стран - членов ЕС.

В соответствии с упомянутыми рекомендациями даже при самом эффективном ведении информационных массивов, содержащих сведения о хозяйствующих на соответствующей территории субъектах, в их составе всегда будет содержаться определенное число «ложно» активных единиц. В связи с этим в целях получения качественного делового регистра предприятий необходимо ограничить процент наличия рассматриваемых объектов до 50% от числа единиц, прекративших деятельность.

Принимая во внимание тот факт, что и по статистическому и по административному разделам Статрегистра данный показатель по Карачаево-Черкесской Республике значительно ниже рекомендованного предела, можно классифицировать уровень качества изучаемого ресурса, который находится в пределах нормы.

Вместе с тем немногим менее трети из числа обследованных объектов не обнаружены по адресу регистрации, что снижает достоверность разрабатываемой на основе Статрегистра информации и создает значительные проблемы в работе по обеспечению полноты охвата предприятий и организаций республики соответствующей статистической и финансовой отчетностью. Кроме того, наличие в Статрегистре таких единиц повлечет за собой неоправданные материальные и временные затраты на их поиск во время проведения экономической переписи.

В связи с этим особую актуальность приобретает вопрос о необходимости внесения изменений в Закон о государственной регистрации юридических лиц и индивидуальных предпринимателей, предусматривающих установление такого порядка, когда достоверность адресных данных и контактных линий, заявленных о себе объектами регистрации, контролируется непосредственно при их включении в ЕГРЮЛ (ЕГРИП), а также порядка, регламентирующего обязательность внесения в определенные сроки изменений к рассматриваемым реквизитам при смене адреса местонахождения предприятий и места жительства индивидуального предпринимателя.

Кроме того, большое значение имеет усиление межведомственного взаимодействия, в развитие которого распоряжением Правительства РФ № 201-р от 14.02.2009 предусматривается представление налоговыми службами в Росстат к 1 сентября 2010 г. перечня сдающих налоговую отчетность налогоплательщиков (как юридических лиц, так и индивидуальных предпринимателей).

Работа в данном направлении, предусматривающая расширение состава ведомств, информационные ресурсы которых можно использовать для актуализации сведений, содержащихся в Статрегистре, а также установление такого взаимодействия не на единовременной, а на периодической основе даст возможность проводить работы по сплошным обследованиям эффективно и с наименьшими затратами, что в конечном счете позволит получить полную информацию о межотраслевых связях и структурных пропорциях российской экономики и создать народнохозяйственный баланс.

УТОЧНЕНИЕ ВЫБОРОЧНЫХ ИТОГОВ С ПОМОЩЬЮ ДОПОЛНИТЕЛЬНЫХ ДАННЫХ. КАЛИБРОВКА ВЫБОРКИ

С.В. Степанов, *канд. социол. наук,*
консалтинговая компания «Планова-Консалтинг»

В статье рассматриваются подходы к обработке данных выборочного наблюдения с учетом дополнительных источников информации на основе использования новых методов выборочного наблюдения.

Решение задачи повышения эффективности выборочных обследований предполагает объединение в комплексном подходе решения разнородных подзадач, включая административные, например:

1. Контроль актуальности (выборочной основы);
2. Повышение уровня квалификации исполнителей выборочных обследований;
3. Совершенствование алгоритмов построения выборочных планов (дизайн выборки) с регулируемыми параметрами;
4. Повышение адекватности и точности выборочных оценок;
5. Снижение нагрузки на респондентов без снижения точности выборочных оценок;
6. Коррекция негативного влияния неответов респондентов и т. п.

Содержание настоящей работы сосредоточено на рекомендациях по решению подзадач 3, 4 и 5 в рамках современных возможностей технической и технологической базы Росстата. Выборочная технология, реализованная в Росстате для выборочных обследований предприятий, достаточно детально разработана и подробно описана в методологическом сборнике[1]. Последовательному продолжению развития статистической методологии в области выборочных обследований и посвящена эта статья в части использования дополнительных, вспомогательных источников информации для оптимизации выборки.

[1] См.: Методологические положения по статистике. Вып. 3. / Госкомстат России. - М., 2000. С. 9-26.

Калибровкой выборки называется процесс целенаправленного изменения таких параметров выборочного плана, как выборочный вес, а также непараметрические модификации состава выборки для сокращения ошибок выборки и повышения точности и устойчивости выборочных оценок статистических показателей. Такое преобразование базового выборочного плана становится возможным при использовании информации из дополнительных источников, например при наличии количественных признаков в Статистическом регистре предприятий, которые характеризуют разные аспекты размера объекта наблюдения, или при использовании данных полных или экономических переписей.

Наличие в нашей стране большого количества регионов с существенно различной степенью экономического развития и с различной экономической структурой делает необходимым использование в статистической практике таких методов, которые не зависели бы от местной конкретной специфики исследуемой совокупности. Методы оценки дисперсии, доверительных интервалов, коэффициента вариации и других точечных и интервальных параметров исследуемой совокупности, основанные на модели (model-based), обладают одной особенностью. Либо они основываются на предположении о намеренном упрощении процедуры отбора и чреваты смещением, связанным с таким упрощением, либо для каждой функции оценивания необходимо находить свою не смещенную формулу, учитывающую особенности конкретной процедуры, что не всегда просто[2]. Независимость от конкретных условий, неизвестных заранее функций распределения единиц наблюдения способны обеспечить непараметрические методы многомерного статистического анализа[3], такие, как методы jackknife (складного ножа) и bootstrap (бутстреп), которые, в частности, требуют значительных объемов вычислений.

1. Анализ современной теории калибровки выборочных весов с использованием дополнительной информации

1.1. Определения калибровки

Метод калибровки выборочных весов получил широкое распространение в службах государственной статистики крупнейших развитых стран и в практике Евростата. Развитие статистической теории, связанной с выборочными оценками и их улучшением путем калибровки выборок, сильно продвинулось с момента публикации в Журнале американской Статистической Ассоциации в 1992 г. теперь широко известной статьи Жан-Клода Девиля[4] и Карла-Эрика Сендала[5] «Оценочные функции калибровки в выборочном обследовании»[6]. [Сам термин «калибровка» происходит от французского слова «calage» («втискивание») и имеет коннотацию[7] «стабильность».]

Основная идея получения калиброванных статистических оценок, предложенная Deville и Särndal, заключается в расчете калиброванных весов выборки при условии ограничения вида: сумма весов первоначального плана выборки и сумма калиброванных весов равны. Дополнительной информацией в виде суммы может являться не только сумма весов, но и, в общем случае, суммы по вспомогательным переменным. Обобщая вышесказанное, можно утверждать, что *методика калибровки* для конечных совокупностей состоит из:

(a) вычисления весов, которые включают определенную *вспомогательную информацию* и ограничены *уравнениями калибровки*;

(b) использования этих весов, чтобы вычислить линейно взвешенные оценки сумм или других параметров конечной совокупности: взвешенных значений переменных, суммированных по определенным группам наблюдаемых объектов;

(c) цели получить *почти не смещенные* выборочные оценки при условии отсутствия неответов и ошибок, не связанных с выборкой.

В литературе термин «калибровка» нередко соотносят только с (a), однако в этих рекомендациях мы часто будем подразумевать (a) вместе с (c). Более ранние определения, по существу, соответствуют приведенным выше. Pascal Ardilly[8] определяет калибровку как метод перевзвешивания, при котором у исследователя есть доступ к нескольким переменным, качественным или количественным, на значениях которых он желает выполнить совместную настройку.

Kott (2006)[9] определяет веса калибровки как ряд весов для объектов в выборке, которые удовлетворяют калибровке к известным итогам по совокупности

[2] См.: Roberts G., Binder D., Kovacevic M., Pantel M., Phillips O. Using an estimating function bootstrap approach for obtaining variance estimates when modelling complex health survey data / SSC Annual Meeting, June 2003. Proceedings of the Survey Methods Section.

[3] См.: Эфрон Б. Нетрадиционные методы многомерного статистического анализа: Сб. статей: Пер с англ. - М.: Финансы и статистика, 1988. С. 19-48.

[4] Жан-Клод Девиль (Jean-Claude Deville) - руководитель отдела статистической методологии и выборки Национального института статистики и экономических исследований Франции [Institut National de la Statistique et des'Etudes'Economiques (INSEE)].

[5] Карл-Эрик Сендал (Carl-Erik Sarndal) - профессор отдела математики и статистики Университета Монреаля, Канада.

[6] См.: Deville Jean-Claude, Sarndal Carl-Erik. Calibration Estimators in Survey Sampling/ Journal of the American Statistical Association. Vol. 87. №. 418. Jun 1992, 376-382.

[7] Коннотация - тип лексической информации, сопутствующей значению слова. Иногда называется также (семантической) ассоциацией. Коннотация слова отражает такой признак обозначаемого им объекта, который хотя и не составляет необходимого условия для применения данного слова, но устойчиво связан с обозначаемым объектом в сознании носителей языка.

[8] См.: Ardilly P. Les techniques de sondage. 2006, Paris: Editions Technip.

[9] См.: Kott P.S. Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology, 2006, 32, 133-142.

и, таким образом, получающаяся оценка не противоречит вероятностному характеру выборочного плана, или, более строго, что смещение выборочного плана, при умеренных условиях, вносит *асимптотически несущественный* вклад в средний квадрат ошибки оценки статистического параметра. Это свойство калибровки Särndal называет «*почти не смещенным* планом выборки».

В четвертом выпуске методологических рекомендаций статистики Канады [The Quality Guidelines (fourth edition) of Statistics Canada (2003)] говорится: «Калибровка - процедура, которую можно использовать, чтобы включить информацию, содержащуюся во вспомогательных данных. Эта процедура корректирует выборочные веса с помощью множителей, известных как факторы (коэффициенты) калибровки, которые приводят оценки статистических параметров в согласие с известными итогами. Получающиеся веса называются весами калибровки или конечными весами оценивания. Эти веса калибровки в общем случае приводят к оценкам, которые не противоречат выборочному плану, и эти оценки имеют меньшую дисперсию, чем оценка Горвица-Томпсона».

Утверждение (c) требует комментария. Ничто не мешает рассчитывать веса, калиброванные к имеющейся вспомогательной информации без учета требования (c). Когда ошибки, не связанные с выборочным планом, присутствуют, смещение в оценках неизбежно, сделаны ли они с применением калибровки или каким-нибудь другим методом. В соответствии с (c) тема предлагаемых методологических рекомендаций ограничивается выборочным планом.

1.2. Развитие применяемых методик взвешивания

Калибровка как линейный метод взвешивания. У калибровки есть близкая, по существу, методика, применяемая на практике. Назначение фиксированных весов, как метод, в практике ведущих национальных статистических агентств был важным и популярным инструментом практики до калибровочных методов.

Назначать соответствующий (вмененный) вес на наблюдаемое значение переменной и суммировать взвешенные значения переменных для того, чтобы сформировать соответствующие сводные показатели - постоянно используемая процедура. Этот способ применяется статистическими службами для оценивания различных описательных параметров конечной совокупности: суммы, средние и функции от сумм. Взвешивание просто объяснить пользователям статистической информации и другим контрагентам статистических служб.

Взвешивание наблюдаемых значений объектов инверсией их вероятности включения нашло твердую научную поддержку в статьях Hansen и Hurwitz (1943)[10], Horwitz и Tompson (1952)[11]. Взвешивание стало широко принято. Позже постстратификационное взвешивание достигло того же уровня популярности. Калибровочное взвешивание обобщает и расширяет обе эти идеи. Калибровочное взвешивание - производное результата, так как веса зависят от *наблюдаемой* выборки.

Веса, обратные вероятности включения, по определению, больше или равны единице. В то же время калиброванные веса необязательно больше или равны единице, если это специально не предусмотрено вычислительным алгоритмом.

Взвешивание значений наблюдаемых переменных было важной темой, прежде чем калибровка стала популярным методом в среде статистических служб. Некоторые авторы получали веса, аргументируя тем, что они должны как можно меньше отличаться от не смещенных весов выборочного плана (обратных вероятности включения). Другие находили веса, подразумевая, что линейная регрессия оценки должна быть записана как линейно взвешенная сумма наблюдаемых значений переменной исследования. Использовались такие термины, как «взвешивание выборочного обследования», «взвешивание по регрессии» и «взвешивание наблюдения». Среди «ранних статей» отметим работы таких авторов, как Alexander (1987), Bankier, Rathwell и Majkowski (1992), Bethlehem и Keller (1987), Chambers (1996), Fuller, Loughin и Baker (1994), Kalton и FloresCervantes (1998), Lemaitre и Dufour (1987)[12], Särndal (1982) и Zieschang (1990)[13]. Более поздний термин «калибровка» имеет более определенный смысл и более точное руководство по производимым вычислениям, чем более старый термин «взвешивание».

Калибровка как системный метод использования вспомогательной информации. Калибровка обеспечивает систематический способ принять во внимание вспомогательную информацию. Как указывают Rueda, Martinez, Martinez и Arcos (2007)[14], «во многих стандартных условиях калибрация обеспечивает простой практический подход к включению вспомогательной информации в оценку».

[10] См.: Hansen M.H., and Hurwitz W.N. On the theory of sampling from finite populations. Annals of Mathematical Statistics, 1943, 14, 333-362.

[11] См.: Horvitz D.G. and Thompson D.J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 1952, 47, 663-685.

[12] См.: Lemaitre G. and Dufour J. An integrated method for weighting persons and families. Survey Methodology, 1987, 13, 199-207.

[13] См.: Zieschang K.D. Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Journal of the American Statistical Association, 1990, 85, 986-1001.

[14] См.: Rueda M., Martinez S., Martinez H. and Arcos A. Estimation of the distribution function with calibration methods. Journal of Statistical Planning and Inference, 2007, 137, 435-448.

Вспомогательная информация использовалась для того, чтобы улучшить точность оценок обследования, прежде чем стала широко использоваться калибровка. Сегодня калибровка действительно предлагает систематизированный подход к использованию вспомогательной информации. Например, калибровка может эффективно применяться в обследованиях, где вспомогательная информация существует на разных уровнях. При осуществлении двухступенчатой выборки одна информация может существовать для первой стадии отбора кластеров, связанных с категориальными переменными принадлежности, а другая информация - для второй стадии отбора объектов в кластерах. В обследованиях с предполагаемыми неответами (то есть, по существу, во всех обследованиях) информация может существовать «на уровне совокупности» (известны суммы по совокупности). Другая информация существует «на выборочном уровне» - значения вспомогательных переменных для всех, включенных в выборку, ответивших и не ответивших.

Калибровка для достижения согласованности. Калибровка часто описывается как способ получить «согласованные оценки». (Здесь «согласованные ...» подразумеваются не в отношении к вероятностному способу выборки, а в смысле «совместимые с известными агрегированными данными»). Уравнения калибровки налагают такие ограничения на систему весов, чтобы когда они применяются к вспомогательным переменным, это подтверждало бы их совместимость с известными агрегированными значениями для этих же самых вспомогательных переменных. Желание усилить доверие к конечным оценкам часто упоминается в публикуемой статистической литературе в связи с задаваемыми требованиями по согласованности. Некоторые пользователи статистической информации находят неудовлетворительным то обстоятельство, при котором две или более оценки по одной и той же совокупности оказываются несогласованными.

Оцениваемые суммы, с помощью которых исследуется согласованность, часто называют *управляемыми* суммами. «Управляемые веса» или «калиброванные веса» предлагают улучшенную, более точную оценку. У согласованности, достигаемой посредством калибровки, есть более широкое значение, чем просто более точное согласование с известными вспомогательными суммами совокупности. Согласованность может быть, например, исследована с соответствующими оценками, получаемыми как в текущем обследовании, так и в других обследованиях и иных (административных) источниках.

Согласованность в таблицах оценок, полученных в результате разных обследований, была поводом для разработки метода *повторяемого взвешивания* - методики, разработанной в голландском национальном статистическом агентстве CBS и опубликованной в ряде статей таких авторов, как Renssen и Nieuwenbroek (1997)[15], Renssen, Kroese и Willeboordse (2001)[16], Knottnerus и van Duin (2006)[17]. Цель метода состоит в том, чтобы приспособить пользовательские запросы для производства согласованных выводов в численном виде. В последней упомянутой статье указывается, что повторяемое взвешивание может быть применено как дополнительный шаг калибровки для новой настройки уже калиброванных весов. Конечные веса принимаются согласованными в заданных (или определенных) границах.

Согласованность с известными или предполагаемыми суммами может принести дополнительный выигрыш в улучшении точности (более низкая дисперсия и/или уменьшенное смещение вследствие неответов). Однако в некоторых статьях, особенно создаваемых в статистических агентствах, согласованность для удовлетворения представлений пользователей кажется более обязательным побуждением, нежели перспектива улучшения точности.

Если первичный мотив для применения калибровки связан не столько с обеспечением согласованности с другими статистиками, сколько с уменьшением дисперсии и сокращением смещения, связанного с неответами, то более соответствующим описанием для калибровки является «система сбалансированных весов», а не «система согласованных весов». Цель состоит в том, чтобы балансировка весов отражала результат выборки, содержание ответов обследования и всю доступную информацию.

Калибровка для удобства и прозрачности. Harms и Duchesne (2006)[18] указывают: «Методика калибровки получила широкое распространение в практике, потому что получающиеся оценки просто интерпретировать и мотивировать, а основаны они на весах выборочного плана и естественных ограничениях калибровки». Калибровка на известных итогах кажется типичному пользователю прозрачной и естественной. Пользователи, которые понимают выборочное взвешивание, ценят влияние калибровки, так как уважают управляемые параметры в том смысле, что выборочные веса

[15] См.: Renssen R.H. and Nieuwenbroek N.J. Aligning estimates for common variables in two or more sample surveys. Journal of the American Statistical Association, 1997, 92, 368-374.

[16] См.: Renssen R.H., Kroese A.H. and Willeboordse A.J. Aligning estimates by repeated weighting. Report, Central Bureau of Statistics, 2001, The Netherlands.

[17] См.: Knottnerus P. and van Duin C. Variances in repeated weighting with an application to the Dutch Labour Force Survey. Journal of Official Statistics, 2006, 22, 565-584.

[18] См.: Harms T. and Duchesne P. On calibration estimation for quantiles. Survey Methodology, 2006, 32, 37-52.

были «только немного изменены». Несмещенность только незначительно нарушена. Более простые формы калибровки не вызывают недоверия, так как применяются только «естественные ограничения». Очень ценится еще одно преимущество: во многих областях применения калибровка дает уникальную систему взвешивания, применимую ко всем переменным обследования, многие из которых есть в больших правительственных обследованиях и иных административных источниках.

Калибровка в комбинации с другими терминами. Некоторые авторы используют термин «калибровка» в комбинации с другими терминами, чтобы описать различные направления рассуждений. Вот примеры этого быстрого увеличения терминов: модель-калибровка (Wu и Sitter, 2001)[19], *g*-калибровка (Vanderhoeft, Waeytens и Museux, 2000)[20], гармонизированная калибровка (Webber, Latouche и Rancourt, 2000)[21], высокоуровневая калибровка (Singh, Horn и Yu, 1998), калибровка по регрессии (Demnati и Rao, 2004), нелинейная калибровка (Plikusas, 2006)[22], сверхобобщенная калибровка (Calage super généralisé, Ardilly 2006)[23], модель-калиброванная оценка с помощью нейронной сети и модель-калиброванная оценка локальным полиномом (Montanari и Ranalli, 2003, 2005)[24], модель-калиброванная псевдо-эмпирическая оценка максимального правдоподобия (Wu и Sitter, 2003)[25] и др. Кроме того, калибровка играет существенную роль в косвенных выборочных методах, предложенных Lavalleé (2006)[26]. В несколько ином аспекте, здесь не рассматриваемом, предложены калиброванные вмененные значения для импутации данных (Beaumont 2005)[27] и смещение калибровки [Chambers, Dorfman и Wehrly (1993), Zheng и Little (2003)[28]]. Вышеперечисленные статьи не дают абсолютно полного обзора всех инноваций в сфере калибровки, но только одни их названия указывают на направления, которые были исследованы.

Калибровка как новое направление исследований. Если калибровка представляет «новый подход» с явными отличиями в сравнении с предшествовавшими, то мы должны исследовать такие вопросы, как: Обобщает ли калибровка более ранние теории или подходы? Дает ли калибровка лучшие, более удовлетворительные ответы на важные вопросы в сравнении с ранее исследованными методами?

Практика выборочных обследований сталкивается с такими «неприятностями», как неответы респондентов, ошибки идентификации и ошибки измерений. Верно, что импутация вмененных значений и перевзвешивание для преодоления смещения, вызванного неответами, широко распространены в практике посредством соответствующих методик. Но это, так или иначе, - «отдельные проблемы», которые все еще ждут, чтобы быть более полно описанными во всесторонней, более удовлетворительной теории вывода в выборочных обследованиях. Много работ, посвященных выборочной теории, касаются оценки для предполагаемого идеального обследования, не существующего практически, где неответы и другие ошибки, не связанные с планом отбора, отсутствуют. Исследования свойств калибровки показывают, что она может обеспечить более систематизированный подход к обследованиям даже в присутствии различных ошибок, не связанных с планом выборки.

1.3. Модель калибровки, основанная на выборочном плане

Представим конечную совокупность $\Omega = \{1, 2, \ldots, i, \ldots, N\}$ из N объектов, из которой выполнена вероятностная выборка s ($s \in \Omega$) фиксированного размера n, получена с вероятностью $p(s)$ в соответствии с планом отбора p. Вероятности включения объекта в выборку $\pi_i = \Pr(i \in s)$[29] и $\pi_{ij} = \Pr(i \neq j \in s)$[30] предполагаются строго положительными и известны. Пусть y_i есть значение

[19] См.: Wu C. and Sitter R.R. A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, 2001, 96, 185-193.

[20] См.: Vanderhoeft C., Waeytens E. and Museux J.M. Generalised calibration with SPSS 9.0 for Windows baser. In Enquêtes, Modèles et Applications (Eds. J.J. Droesbeke and L. Lebart), 2001, Paris: Dunod.

[21] См.: Webber M., Latouche M. and Rancourt E. Harmonised calibration of income statistics. Statistics Canada, internal document, April 2000.

[22] См.: Plikusas A. Nonlinear calibration. Proceedings. Workshop on Survey Sampling, 2006, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

[23] См.: Ardilly P. Op. cit.

[24] См.: Montanari G.E. and Ranalli M.G. On calibration methods for design-based finite population inferences. 2006, Bulletin of the International Statistical Institute, 54th session, volume LX, contributed papers, book 2, 81-82; Montanari G.E. and Ranalli M.G. Nonparametric model-calibration estimation in survey sampling. Journal of the American Statistical Association, 2005, 100, 1429-1442.

[25] См.: Wu C. and Sitter R.R. Op. cit.

[26] См.: Lavallée P. Indirect Sampling. 2007, New York: SpringerVerlag.

[27] См.: Beaumont J.F. Calibrated imputation in surveys under a quasi model-assisted approach. Journal of the Royal Statistical Society B, 2005, 67, 445-458.

[28] См.: Zheng H. and Little R.J.A. Penalized spline model-based estimation of the finite population total from probability-proportional-to-size-samples. Journal of Official Statistics, 2003, 19, 99-117.

[29] Вероятность включения первого порядка.

[30] Вероятность включения второго порядка.

интересующей нас переменной y для i-го объекта совокупности, с которым также связана вспомогательная переменная x_i или вектор вспомогательных переменных $\mathbf{x}_i$. Для объекта из выборки $i \in s$ мы можем наблюдать и измерить (y_i, x_i). Сумма по совокупности вспомогательной переменной x, $X=\sum_{i\in\Omega} x_i$ нам доступна и известна. Цель состоит в том, чтобы получить оценку неизвестной нам суммы по совокупности - $\sum_{i\in\Omega} y_i$. Для калибровки, для которой ведутся эти рассуждения, важно точно определить *вспомогательную информацию*. При прочих основных условиях мы должны различать две ситуации относительно $\mathbf{x}_i$:

(i) $\mathbf{x}_i$ - известный вектор значений для каждого $i\in\Omega$ (полная вспомогательная информация);

(ii) $\Sigma\mathbf{x}_i$ - известные (полученные извне) итоги, и $\mathbf{x}_i$ известен (измерен в обследовании) для каждого $i \in s$.

Зачастую среда или обстоятельства обследования диктуют ситуацию (i) или (ii), по преимуществу. Вариант (i), то есть наличие полной вспомогательной информации, имеет место, когда значения вектора $\mathbf{x}_i$ определены и известны по всей выборочной основе для каждого $i \in \Omega$ (и соответственно для каждого $i \in s$). Такая ситуация типична при индивидуальных обследованиях населения и при обследованиях домашних хозяйств, к примеру в Скандинавии и странах Северной Европы, имеющих в своем распоряжении высококачественные административные регистры, которыми можно воспользоваться в качестве основы выборки, чтобы обеспечить большое количество потенциальных вспомогательных переменных. Итоги по совокупности $\sum_{i\in\Omega}\mathbf{x}_i$ можно получить, просто складывая $\mathbf{x}_i$.

Вариант (i) дает значительную свободу по структурированию вспомогательного вектора $\mathbf{x}_i$. К примеру, если x_i являются значениями непрерывной переменной, определенными для каждого $i \in \Omega$, то мы имеем возможность рассмотреть x_i^2 и другие функции от x_i для включения их в $\mathbf{x}_i$, потому что итоги, такие, как $\sum_{i\in\Omega} x_i^2$ и $\sum_{i\in\Omega}\log x_i$, могут быть легко вычислены. Если зависимости с изучаемой переменной нелинейны, то было бы большим упущением не принять во внимание такие доступные формы итогов, как квадратичные или логарифмические.

Вариант (ii) преобладает в обследованиях, где ситуация (i) не встречается, но где $\sum_{i\in\Omega}\mathbf{x}_i$ можно получить из внешних источников, которые считаются достаточно точными, а индивидуальные значения $\mathbf{x}_i$ доступны (измерены в процессе сбора данных) для каждого $i \in s$. В этом случае $\sum_{i\in\Omega}\mathbf{x}_i$ иногда называют «независимым управляющим итогом», чтобы отметить его происхождение, внешнее по отношению к обследованию. Вариант (ii) менее гибок: если x_i является переменной с итогом $\sum_{i\in\Omega} x_i$, взятым из внешнего надежного источника, то $\sum_{i\in\Omega} x_i^2$ может быть недоступна, лишая возможности включения x_i^2 в вектор вспомогательных переменных $\mathbf{x}_i$.

1.4. Обобщенная оценка по регрессии для базовых условий. Концепция GREG

Прежде чем рассматривать различные реализации калибровки, следует определиться в отношении *оценивания с помощью обобщенной регрессии* [*generalized regression (GREG) estimation*] (или точнее, *регрессионное оценивание*) с позиций двух серьезных точек зрения: (1) во многих статьях справедливо утверждается, что GREG-оценивание есть систематизированный способ принять во внимание вспомогательную информацию; (2) некоторые (но не все) GREG-оценки есть оценки калибровки, которые могут быть выражены в терминах (калиброванного) линейного взвешивания. За прошлые два десятилетия интенсивно изучались GREG-оценки и оценки калибровки. Одни только термины «GREG-оценка» и «оценка калибровки» отражают ясное различие в методологическом подходе.

Понятие GREG-оценки постепенно развивалось с середины 1970-х годов. Простая (линейная) GREG-оценка описывается в работе Särndal, Swensson and Wretman (1992)[31]. Центральная идея состоит в том, что предсказанные y-значения $\hat{y}_i$ могут быть вычислены для всех N элементов совокупности с помощью подобранной *вспомогательной модели* и использования вспомогательного вектора значений $\mathbf{x}_i$, известных для каждого $i \in \Omega$. Предсказанные значения служат для того, чтобы построить *почти не смещенную* оценку суммы $Y = \sum_{i\in\Omega} y_i$, зависимую от выборочного плана:

$$\hat{Y}_{GREG} = \sum_{\Omega}\hat{y}_i + \sum_s d_i(y_i - \hat{y}_i) = \tag{1}$$

$$= \sum_s d_i y_i + (\sum_{\Omega}\hat{y}_i - \sum_s d_i\hat{y}_i).$$

Очевидная цель для создания такой конструкции - это перспектива получения очень точной оценки $\hat{Y}_{GREG}$ с помощью тонкого подбора вспомогательной модели, которая дает очень маленькие остатки y_i - $\hat{y}_i$. Такое моделирование - краеугольный камень рассуждений в стиле «GREG-оценка». Некоторые авторы для конструкции (1) используют термин «общая оценка отличия».

[31] См.: Särndal C.E., Swensson B. and Wretman J. Model-assisted Survey Sampling. 1992, New York: SpringerVerlag.

Большое разнообразие возможных вспомогательных моделей порождает большое семейство GREG-оценок формы (1). У вспомогательной модели, предполагающей зависимость между **x** и *y*, может быть много форм: линейная, нелинейная, обобщенная линейная, смешанная (модель с некоторыми фиксированными и некоторыми случайными параметрами) и т. д. Безотносительно выбора формы, модель «только помогает» даже притом, что форма может быть «истинной», (1) соответствует плану выборки, *почти не смещенному* при умеренных условиях для вспомогательной модели и самого плана выборки.

1.5. Линейная GREG-оценка

Под линейной GREG-оценкой мы будем понимать такую, которая получена с помощью вспомогательной модели с линейными коэффициентами. Предсказанные значения $\hat{y}_i = \mathbf{x}_i' \mathbf{B}_{s;dq}$, где

$$\mathbf{B}_{s;dq} = (\sum_s d_i q_i \mathbf{x}_i \mathbf{x}_i')^{-1} (\sum_s d_i q_i \mathbf{x}_i y_i),$$

совместно с (1) дает:

$$\hat{Y}_{GREG} = (\sum_\Omega \mathbf{x}_i)' \mathbf{B}_{s;dq} + \sum_s d_i (y_i - \mathbf{x}_i' \mathbf{B}_{s;dq}).$$

Здесь q_i - масштабные коэффициенты, выбираемые статистиком. Стандартный выбор: $q_i = 1$ для всех *i*. У выбора q_i есть некоторое (но часто ограниченное) влияние на точность $\hat{Y}_{GREG}$; *почти не смещенность* сохраняется для любого выбора q_i, за исключением чрезмерных. Хотя модель проста, используя линейную GREG (2), можно сформулировать много оценок, рассматривая множество возможных выборов вектора вспомогательных переменных $\mathbf{x}_i$ и масштабных коэффициентов q_i. При общих условиях

$$(\hat{Y}_{GREG} - Y) / N = (\sum_s d_i E_i - \sum_\Omega E_i) / N + O_p(n^{-1}),$$

где $\sum_s d_i E_i$ - это оценка Горвица-Томпсона для остатков $E_i = y_i - \mathbf{x}_i' \mathbf{B}_{\Omega;q}$. Следовательно, связанные с планом выборки показатели $E(\hat{Y}_{GREG}) \approx Y$ и $Var(\hat{Y}_{GREG}) \approx Var(\sum_s d_i E_i)$. Близкая подгонка *линейной* регрессии для *y* от **x** является ключом к малой дисперсии $\hat{Y}_{GREG}$ (и это очень отличается от утверждения, что «линейная регрессия - это истинная регрессия»).

Линейная GREG-оценка в работе авторов Särndal, Swensson и Wretman (1992)[32] обосновывалась с помощью вспомогательной линейной модели ξ, формулируемой так: $E_\xi(y_i) = \boldsymbol{\beta}\mathbf{x}_i$ и $V_\xi(y_i) = \sigma_i^2$. Метод обобщенных наименьших квадратов дает оценку (2) с $q_i = 1/\sigma_i^2$. В этом контексте обоснованное предположение о разбросе остатков $y_i - \boldsymbol{\beta}\mathbf{x}_i$ определяет и q_i. Когда вектор $\mathbf{x}_i$ фиксирован, основные усилия по подбору модели сводятся к определению модели остатков. Выбор $\sigma_i^2 = \sigma_i^2 x_i$ дает классическую оценку по отношению. Если $q_i = \boldsymbol{\mu}\mathbf{x}_i$ для всех $i \in \Omega$ и $\boldsymbol{\mu}$ есть вектор констант, то тогда (2) сокращается до формы $(\sum_\Omega \mathbf{x}_i)' \mathbf{B}_{s;dq}$.

Beaumont и Alavi (2004)[33] доказали, что линейная GREG-оценка является устойчивой по отношению к смещению (*почти не смещенной*, хотя вспомогательная модель теряет «корректность»), но она может быть значительно менее эффективна (имеет больший средний квадрат ошибки), чем альтернативные модели, имеющие хоть и большие смещения, но со значительно более малыми дисперсиями. Таким образом, можно утверждать, что линейная GREG-оценка не устойчива по отношению к дисперсии. Это - фундаментальное понятие теории выборочных обследований, основанных на выборочном плане.

Спецификация вектора $\mathbf{x}_i$ должна включать переменные (с известными итогами), которые уже использовались для построения выборочного плана. Информация стадии построения выборочного плана не должна быть исключена на стадии оценивания, напротив, рекомендуется ее «повторное» использование. Например, в варианте простой случайной стратифицированной выборки вектор $\mathbf{x}_i$ в статистической оценке (2) должен включать, наряду с другими доступными переменными, атрибутивные коды принадлежности к страте.

Мы можем записать линейную GREG-оценку (2) как взвешенную выборочную сумму $\hat{Y}_{GREG} = \sum_s w_i y_i$ с весами:

$$w_i = d_i g_i; \quad g_i = 1 + q_i \boldsymbol{\lambda}\mathbf{x}_i; \qquad (3)$$
$$\boldsymbol{\lambda}' = (\sum_\Omega x_i - \sum_s d_i x_i)' (\sum_s d_i q_i \mathbf{x}_i \mathbf{x}_i')^{-1}.$$

Веса w_i *калиброваны* (совместимы с) известными **x**-итогами по совокупности: $\sum_s w_i \mathbf{x}_i = \sum_\Omega \mathbf{x}_i$. То, что $\hat{Y}_{GREG}$ выражен как линейная взвешенная сумма с калиброванными весами, является побочным эффектом формализованного вывода. Это не часть GREG-стиля рассуждений, центральная идея которого, сформулированная в (1), есть подбор вспомогательной модели.

1.6. Калибровка при базовых условиях выборки

Решающий прием в использовании GREG-подхода - это предсказание значений $\hat{y}_i$ с помощью подбора вспо-

[32] См.: Ibid.

[33] См.: Beaumont J.F. and Alavi A. Robust generalized regression estimation. Survey Methodology, 2004, 30, 195-208.

могательной модели. В противоположность этому, калибровочный подход, определенный в п. 1.1, не обращается явно ни к какой модели. Вместо этого подчеркивается существенность использования вспомогательной информации, по которой можно калибровать. Ключевым моментом в рассуждениях в калибровочном духе является линейное взвешивание наблюдаемых *y*-значений с весами, согласованными с вычисляемыми итогами. Это концептуальное различие будет приводить иногда к различным статистическим оценкам в этих двух подходах.

Калибровочный подход обладает значительной степенью общности - он может быть применен на множестве условий: сложные выборочные планы, корректировки неответов и ошибки основы выборки. Остановимся, однако, на базовых условиях: единственная стадия выборки и полные ответы. Данные, доступные для того, чтобы получить оценки итогов по совокупности $Y = \sum_{i\in\Omega} y_i$, таковы:

1) значения исследуемой переменной y_i получены в наблюдении для $i \in s$;

2) известны веса выборочного плана $d_i = 1/\pi_i$ для $i \in \Omega$;

3) известен вектор переменных $\mathbf{x}_i$ для $i \in \Omega$ или итоги, полученные из внешних источников $\sum_{i\in\Omega} \mathbf{x}_i$.

Эти простые условия преобладают в описаниях Deville и Särndal (1992) и Deville, Särndal и Sautory (1993), статьях, которые дали подходу название и инициировали дальнейшую исследовательскую работу по калибровке. Даже притом, что условия ситуации просты, калибровка поднимает несколько проблем, некоторые из которых чисто вычислительные.

Цель методов калибровки в том, чтобы определить веса w_i, которые удовлетворяли бы уравнениям ограничений калибровки $\sum_s w_i \mathbf{x}_i = \sum_\Omega \mathbf{x}_i$, затем использовать их в формуле калиброванной оценки для Y в виде: $\hat{Y}_{CAL} = \sum_s w_i y_i$, которую мы можем сопоставить с не смещенной оценкой Горвица-Томпсона, записав ее так: $\hat{Y}_{CAL} = \hat{Y}_{HT} + \sum_s (w_i - d_i) y_i$. Из этого следует, что смещение Y_{CAL} есть $E(\hat{Y}_{CAL}) - Y = E(\sum_s (w_i - d_i) y_i)$. Цель достижения п*очти не смещенного* плана выборки требует $E(\sum_s (w_i - d_i) y_i) \approx 0$ независимо и безотносительно *y*-переменной. Очевидно, калибровка должна способствовать уменьшению отклонений $(w_i - d_i)$.

Цель «калибровки, приводящей в соответствие с известными вспомогательными итогами по совокупности», может быть реализована многими путями. Мы можем получить множество наборов калиброванных весов для известных $\sum_\Omega \mathbf{x}_i$. Мы остановимся на методе минимизации расстояния и методе инструментального вектора. Некоторые другие методы конструирования калиброванных весов предложены в работе авторов Demnati и Rao (2004)[34].

1.7. Метод минимального расстояния

В этом методе калибровкой предполагается изменять начальные веса $d_i = 1/\pi$ на новые веса w_i, определенные как «близкие» к d_i. В связи с этим можно ввести функцию расстояния $G_i(w, d)$, определенную для каждого $w > 0$, такого, что $G_i(w, d) \geq 0$, $G_i(d, d) = 0$, дифференцируемую относительно w, строго выпуклую, с непрерывной производной $g_i(w, d) = \partial G_i(w,d)/\partial w$, такой, что $g_i(d, d) = 0$. Обычно функция расстояния выбирается таким образом, что $g_i(w, d) = g_i(w/d)/q_i$, где q_i - соответственно выбранные масштабные коэффициенты, $g(\cdot)$ - функция одного аргумента, непрерывная, строго возрастающая, с $g(1) = 0$, $g'(1) = 1$. Пусть $F(u) = g^{-1}(u)$ - функция, обратная $g(\cdot)$. Минимизируя сумму расстояний $\sum_s G_i(w_i d_i)$ на ограничениях уравнений калибровки $\sum_s w_i \mathbf{x}_i = \sum_\Omega \mathbf{x}_i$, получаем $w_i = d_i F(q_i \mathbf{x}_i \boldsymbol{\lambda})$, где $\boldsymbol{\lambda}$ получен как решение (предполагаем, что оно существует) для:

$$\sum_s d_i \mathbf{x}_i F(q_i \mathbf{x}_i' \boldsymbol{\lambda}) = \sum_\Omega \mathbf{x}_i . \qquad (4)$$

У весов есть свойство оптимальности, так как целевая функция минимизируется должным образом, но это - «слабая оптимальность» в том смысле, что есть много возможных форм функции расстояния и масштабных коэффициентов q_i.

Большое внимание уделяется функции расстояния вида

$$G_i(w_i, d_i) = (w_i - d_i)^2 / 2 d_i q_i .$$

Она дает $g_i(w_i, d_i) = (w_i / d_i - 1) / q_i$; $g(w / d) = w / d - 1$; $F(u) = g^{-1}(u) = 1 + u$. Термин «линейный вариант» в этом случае вполне подходит. Задача состоит в том, чтобы минимизировать «хи-квадрат расстояние» $\sum_s (w_i - d_i)^2 / 2 d_i q_i$ на ограничениях $\sum_s w_i \mathbf{x}_i = \sum_\Omega \mathbf{x}_i$. Выра-

[34] См.: Demnati A. and Rao J.N.K. Linearization variance estimators for survey data. Survey Methodology, 2004, 30, 17-26.

жение (4) запишем как $\sum_s d_i \mathbf{x_i}(1+q_i \mathbf{x_i}'\lambda) = \sum_\Omega \mathbf{x_i}$, чтобы проще вычислить $\boldsymbol{\lambda}$. Получающаяся оценка для $Y = \sum_{i\in\Omega} y_i$ есть $\hat{Y}_{CAL} = \sum_s w_i y_i$ с весами $w_i = d_i g_i$, данными в (3). Поэтому $\hat{Y}_{CAL} = \hat{Y}_{GREG}$, как указано в (2), и остатки, которые определяют асимптотическую дисперсию $E_i = y_i - \mathbf{x_i}'\mathbf{B}_{\Omega:q}$, как дано в п. 1.5. Имеется возможность получения некоторого количества отрицательных весов w_i.

Линейная GREG-оценка подразумевает веса, которые как будто калиброваны (по $\sum_{i\in\Omega} \mathbf{x_i}$), с другой стороны, линейный случай калибровки (с функцией расстояния по хи-квадрат) приводит к линейной GREG-оценке.

Уравнение калибровки удовлетворяется для любого выбора неотрицательных масштабных коэффициентов q_i в (4). Простой выбор - $q_i = 1$ для всех i, но это не всегда предпочтительный выбор. К примеру, если есть единственная, всегда неотрицательная вспомогательная переменная $\mathbf{x}_i = x_i$, то многие будут интуитивно предполагать, что $\hat{Y}_{CAL} = \sum_s w_i y_i$ приводит к обычной оценке по отношению: $\sum_\Omega x_i \frac{\sum_s d_i y_i}{\sum_s d_i x_i}$, и так оно и есть, но при $q_i = x_i^{-1}$, а не при $q_i = 1$.

Значительный интерес представляет другая функция расстояния: $G_i(w_i, d_i) = \{w_i \log(w_i / d_i) - w_i + d_i\} / d_i$. Из нее следует $F(u) = g^{-1}(u)$ - $\exp(u)$, «экспоненциальный случай». Тогда (4) запишем как $\sum_s d_i \mathbf{x_i} \exp(q_i \mathbf{x_i}'\lambda) = \sum_\Omega \mathbf{x_i}$. Численные методы требуют решения по $\boldsymbol{\lambda}$ для получения весов $w_i = d_i \exp(q_i \mathbf{x_i} \boldsymbol{\lambda})$. И никаких отрицательных весов не возникнет.

Deville и Särndal (1992)[35] показали, что множество функций расстояния, удовлетворяющих умеренным условиям, производят асимптотически эквивалентные статистические оценки калибровки. Альтернативные функции калибровки рассмотрены и подвергнуты сравнению в работах авторов Deville, Särndal и Sautory[36] (1993), Singh и Mohl (1996), Stukel, Hidiroglou и Särndal (1996)[37]. Некоторые функции расстояния можно задать так, чтобы они гарантировали, что веса будут находиться в пределах указанных границ для того, чтобы исключить возможность появления слишком больших или слишком маленьких (отрицательных) весов. Модификации функции расстояния часто имеют только незначительное влияние на дисперсию калиброванной оценки $\hat{Y}_{CAL} = \sum_s w_i y_i$, если даже объем выборки будет довольно малым.

1.8. Вычислительные проблемы, критические веса и выбросы

Вычисление калиброванных весов поднимает важные вычислительные проблемы, обсуждаемые во многих статьях. Все вычисления должны проходить гладко и без чрезмерного вмешательства в практику получения информационного продукта национальным статистическим агентством. Многие практики придерживаются разумного требования, чтобы все калиброванные веса были позитивными (больше или равными единице), а также требования избегать очень больших весов. Некоторые из весов, согласно формулам их вычисления, могут оказаться очень большими или отрицательными. Park и Fuller (2005)[38] предлагают методы, позволяющие избежать нежелательных весов. В методе минимизации расстояния функция расстояния может быть сформулирована так, чтобы отрицательные веса были исключены, но при этом удовлетворяли попрежнему уравнениям ограничений калибровки.

Программа CALMAR (Deville, Särndal и Sautory, 1993)[39], применяемая во Франции, позволяет задавать несколько функций расстояния этого вида. Другие статистические агентства разработали свое собственное программное обеспечение для вычисления весов. Среди них - GES (статистическая служба Канады), CLAN97 (статистическая служба Швеции), Bascula 4.0 (Центральное бюро статистики, Нидерланды), g-CALIB (статистическая служба Бельгии). Все они имеют различные средства для преодоления вычислительных проблем.

GES использует линейное программирование для минимизации функции расстояния вида хи-квадрат на ограничениях калибровки по индивидуальным границам весов. Программа g-CALIB, описанная в работах авторов Vanderhoeft, Waeytens и Museux (2001)[40], использует обобщенную инверсию матриц по Муру-Пен-

[35] См.: Deville Jean-Claude., Särndal Carl-Erik. Op. cit.

[36] См.: Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993) Generalized Raking Procedures in Survey Sampling, Journal of the American Statistical Association, Vol. 88, No. 423, 1013-1020.

[37] См.: Stukel D.M., Hidiroglou M.A. and Särndal C.E. Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. Survey Methodology, 1996, 22, 117-125.

[38] Park M. and Fuller W.A. Towards nonnegative regression weights for survey samples. Survey Methodology, 2005, 31, 85-93.

[39] Deville J.-C., Särndal C.-E. and Sautory O. (1993). Op.cit.

[40] См.: Vanderhoeft C., Waeytens E. and Museux J.M. Generalised calibration with SPSS 9.0 for Windows baser. In Enquêtes, Modèles et Applications (Eds. J.J. Droesbeke and L. Lebart), 2001, Paris: Dunod.

роузу для вычисления веса, следовательно, это предотвращает возможное вырождение матриц вследствие избыточности (и линейной зависимости) вспомогательной информации.

Вмешательство в ограничение весов поднимает вопрос о том, как далеко можно отклониться от весов выборочного плана d_i, не ставя под угрозу цель получения плана, дающего *почти не смещенные* оценки. Идея состоит в том, чтобы изменить набор ограничений так, чтобы не нарушались разрешенные допуски между статистической оценкой вспомогательных переменных и соответствующими известными итогами. К примеру, Chambers и Dorfman (1996)[41] предлагают минимизировать «функцию потерь с быстро возрастающей стоимостью».

Значения выбросов во вспомогательных переменных могут быть причиной критических весов. Калибровка при наличии выбросов обсуждается в работе автора Duchesne (1999)[42]. Его методика «робастной калибровки» может добавить определенное смещение в оценках, но это, однако, может быть более чем компенсировано сокращением дисперсии. В тех случаях, когда набор ограничений расширяется ограничениями с допустимыми интервалами для весов, решение задачи оптимизации не гарантируется. Вопросы существования решения рассматриваются в работе автора Thèberge (2000)[43], там же предлагаются методы для ситуаций с выбросами.

2. О возможности использования предлагаемой методики в практике Росстата

2.1. Вычисление калиброванных весов в случае только выборочной вспомогательной информации

Две особенности линейной GREG-оценки (2) делают ее популярным инструментом в практике статистических служб:

1) итоги по совокупности вспомогательных переменных $\sum_{\Omega}\mathbf{x}_i$ могут быть «вынесены за скобки», и процесс вычисления оценки может продолжаться так долго, пока точные значения итогов не будут вычислены или получены из внешних источников;

2) хотя оценка и записана как линейная взвешенная сумма $\hat{Y}_{GREG}=\sum_s w_i y_i$, система весов (3) независима от конкретной y-переменной и, таким образом, может быть применена ко всем y-переменным в обследовании.

Нам не нужно знать значения вектора $\mathbf{x}_i$ для каждого объекта $i\in\Omega$ по всей совокупности, знания итогов $\sum_{\Omega}\mathbf{x}_i$ достаточно. Само собой разумеется, если мы знаем все $\mathbf{x}_i$ для $i\in\Omega$, могут быть найдены более эффективные члены семейства GREG-оценок (все еще *почти не смещенные*). Это можно противопоставить другой критике линейных GREG-оценок, а именно о том, что линейная модель не реалистична для многих моделей данных. Например, для дихотомической y-переменной логарифмическая вспомогательная модель может быть и более реалистичной и может привести к более точной (нелинейной) GREG-оценке.

Мы можем суммировать анализ GREG-оценки для калибровки весов следующим образом. У линейной GREG-оценки есть практические преимущества для масштабного применения в практике выборочных наблюдений Росстата. Такая оценка может быть выражена как линейная взвешенная сумма значений интересующей нас переменной с весами, калиброванными по известным итогам $\sum_{\Omega}\mathbf{x}_i$. Веса независимы от значений y-переменной и могут быть применены ко всем y-переменным в обследовании. Для этого достаточно знать значения вспомогательных итогов $\sum_{\Omega}\mathbf{x}_i$, полученных из надежного источника. Нелинейная GREG-оценка может дать значительно уменьшенную дисперсию в результате применения более совершенных моделей, которые можно рассматривать, когда есть *полная* вспомогательная информация (известен вектор $\mathbf{x}_i$ для каждого объекта $i\in\Omega$), *почти не смещенность* выборочного плана сохраняется. Определенные нелинейные GREG-оценки могут быть сформулированы как линейно взвешенные суммы.

В академических упражнениях с искусственно созданными совокупностями и зависимостями можно вызвать ситуации, где у нелинейной GREG-оценки есть большое преимущество в величине дисперсии перед линейной GREG-оценкой. Такие эксперименты важны для иллюстрации. Однако при рассмотрении ежедневных практических потребностей Росстата в организации выборочных обследований и обработке их данных «неправдоподобная» форма нелинейной GREG-оценки, кажется, представляет в данный момент довольно отдаленный интерес. Вспомогательные модели для GREG-оценки должны отвечать требованиям надежности и практичности, в том числе иметь хорошо объясняемую форму, связанную с социально-экономическими смыслами. Привлекательность незначительного сокращения дисперсии может быть уничтожена проблемами в другом - ошибками, не связанными с мо-

[41] См.: Chambers R.L., Dorfman A.H. and Wehrly T.E. Bias robust estimation in finite populations nonparametric calibration. Journal of the American Statistical Association, 1993, 88, 268-277.

[42] См.: Duchesne P. Robust calibration estimators. Survey Methodology, 1999, 25, 43-56.

[43] См.: Thèberge A. Calibration and restricted weights. Survey Methodology, 2000, 26, 99-107.

делированием и выборочным планом, ошибками наблюдения и прочими неприятностями, происходящими в ежедневном рабочем процессе статистической службы. Прогресс методологии от линейной к нелинейной GREG-оценке создает новые возможности, но и порождает вопросы. Какова самая соответствующая формулировка математического ожидания для нелинейной оценки? Насколько чувствительны результаты к спецификации вспомогательной модели в части формулы дисперсии? До какой степени является проблемой вычислительная эффективность в широком понимании этого термина?

Deville и Särndal (1992)[44] предложили калиброванную оценку:

$$\hat{Y}_{ds} = \sum_{i \in s} w_i y_i \qquad (5)$$

для оценки Горвица-Томпсона (1952)[45]:

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i , \qquad (6)$$

где $d_i = 1/\pi_i$ - базовый вес плана выборки и калиброванные веса w_i, $i \in s$, полученные при минимизации функции расстояния типа хи-квадрат:

$$\sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i} \qquad (7)$$

на ограничениях калибровки:

$$\sum_{i \in s} w_i x_i = X . \qquad (8)$$

Здесь q_i, $i \in s$ - это неотрицательные константы, с помощью которых можно задать веса объектов, не связанные с выборочным планом. В большинстве ситуаций значение q_i принимают равным 1. Форма статистической оценки (5) зависит от выбора q_i. Минимизация (7), подчиненная уравнению калибровки (8), приводит к калиброванным весам вида:

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i \in s} d_i q_i x_i^2} (X - \sum_{i \in s} d_i x_i) . \qquad (9)$$

Подстановка значений w_i из (9) в (5) приводит к обобщенной регрессионной оценке (GREG) суммы Y по всей совокупности:

$$\hat{Y}_{GREG} = \sum_{i \in s} d_i y_i + \hat{\beta}_{ds} (X - \sum_{i \in s} d_i x_i) ,$$

где

$$\hat{\beta}_{ds} = \frac{\sum_{i \in s} d_i q_i x_i y_i}{\sum_{i \in s} d_i q_i x_i^2} .$$

Выражения в предложенной форме вполне пригодны для включения в расчетный алгоритм для расчета калиброванных весов и вычисления калиброванных оценок итогов для простого случая с одной вспомогательной переменной или функцией, объединяющей несколько вспомогательных переменных.

2.2. Пример калибровки по данным Республики Коми

Была обследована совокупность предприятий в Республике Коми. По переписи насчитывалось 3910 предприятий с ненулевой выручкой.

Составим выборочный план, выполним выборку, подставим данные обследования и вычислим оценки в среде SPSS.

Таблица 1

Описательные статистики

	N	Минимум	Максимум	Сумма	Среднее	Дисперсия
Выручка	3910	0	347404,0	11448544,8	2928,017	140696220,1
Численность	4123	0	100	53217	12,91	289,176
N валидных (целиком)	3910					

Представленные гистограммы (см. рис. 1 и 2) показывают, что распределение предприятий как по величине *выручки*, так и по *численности работников* сильно отличается от нормального распределения.

[44] См.: Deville Jean-Claude., Sarndal Carl-Erik. Op. cit.

[45] См.: Horvitz D.G. and Thompson D.J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 1952, 47, 663-685.

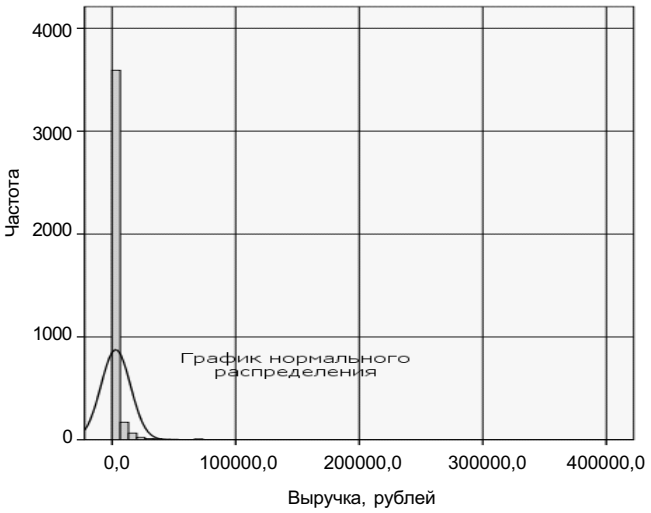

Рис. 1. **Гистограмма распределения предприятий по размеру выручки**

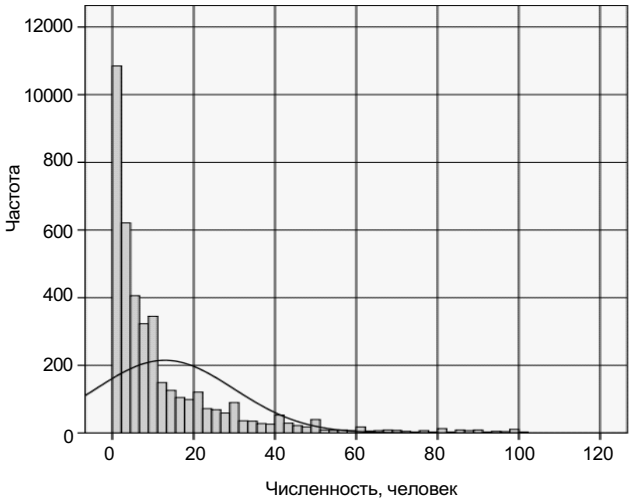

Рис. 2. **Гистограмма распределения предприятий по численности работников**

Основа выборки - Статистический регистр предприятий. Выборочный план: простая случайная выборка без возвращения, объем выборки - 20%, всего выбрано 981 предприятие.

Таблица 2

Оценки Горвица-Томпсона по выборке

		Оценка	Стандартная ошибка	95%-ный доверительный интервал (границы)		Коэффициент вариации	Эффект плана
				Нижняя	Верхняя		
Среднее	Выручка	3068,770	338,0301	2405,248	3732,292	0,110	1,000
	Численность	12,68	0,521	11,66	13,71	0,041	1,000
Сумма	Выручка	12392774,8	1,3651 E6	9713239,3	15072310,4	0,110	1,000
	Численность	54008	2219,169	49652	58364	0,041	1,000

Изучаемая y-переменная - *выручка*. Вспомогательная x-переменная - *численность*, ее известный итог - 53217, оценка 54008.

Вспомогательная переменная одна, поэтому масштабные коэффициенты q_i можно принять равными единице.

Калиброванные веса согласно (9):

$$w_i = 5{,}0 + \frac{5{,}0 \times x_i}{5{,}0 \times 383226}(53217 - 54008).$$

Таблица 3

Оценки по выборке с калиброванными весами

		Оценка	Стандартная ошибка	95% -ный доверительный интервал (границы)		Коэффициент вариации
				Нижняя	Верхняя	
Среднее	Выручка	3046,753	332,9601	2393,183	3700,323	0,109
	Численность	12,56	0,512	11,56	13,57	0,041
Сумма	Выручка	12242467,3	1,3371 E6	9617833,9	14867100,6	0,109
	Численность	53239	2157,010	49005	57473	0,041

Сведем полученные результаты в отдельную таблицу. Эффект, достигаемый калибровкой, будем оценивать по относительному смещению в процентах [Relative Percentage Bias (RB%)] в сравнении с истинной суммой и оценкой Горвица-Томпсона, полученной по выборке с начальными весами: $RB_{HT}(\%) = \frac{\hat{Y}_{HT} - \sum_{i=1}^{3910} y_i}{\sum_{i=1}^{3910} y_i}$, где $\hat{Y}_{HT}$ - оценка Горвица-Томпсона; $\sum_{i=1}^{3910} y_i$ - истинная сумма интересующей нас переменной, полученная по сплошному обследованию. Аналогичная формула применена для калиброванной оценки $\hat{Y}_{CAL}$.

Таблица 4

Эффект калибровки

	Истинная сумма по сплошному обследованию	Оценка Горвица-Томпсона с начальными весами		Оценка с калиброванными весами	
	$\sum_{i=1}^{3910} y_i$	Сумма $\hat{Y}_{HT}$	Относительная ошибка RB_{HT} %	Сумма $\hat{Y}_{CAL}$	Относительная ошибка RB_{CAL} %
Выручка	11448544,8	12392774,8	8,25	12242467,3	6,93
Численность	53217	54008	1,48	53239	0,04

Резкое снижение ошибки по калиброванной оценке *численности* связано с тем, что именно эта переменная была нами выбрана в качестве вспомогательной и именно по ее итогу проводилась калибровка. Очевидное снижение ошибки по *выручке* не столь заметно в предлагаемом случае, так как и некалиброванная оценка уже достаточно хороша. Тем не менее даже в таком варианте калибровка дает ощутимый эффект снижения смещения оценки, и этот эффект будет тем больше, чем хуже будет некалиброванная оценка и чем больше ошибок, не связанных с выборочным планом (неответы, недостаточная актуальность основы выборки и пр.), встретится в процессе проведения обследования.

ОБЪЯВЛЕНИЕ
о конкурсе на замещение вакантных должностей федеральной государственной гражданской службы в Федеральной службе государственной статистики

Федеральная служба государственной статистики (далее - Росстат) информирует о том, что в соответствии с Федеральным законом от 27 июля 2004 г. № 79-ФЗ «О государственной гражданской службе Российской Федерации», Указом Президента Российской Федерации от 1 февраля 2005 г. № 112 «О конкурсе на замещение вакантной должности государственной гражданской службы Российской Федерации», Методикой проведения конкурса на замещение вакантной должности федеральной государственной гражданской службы в Росстате, утвержденной приказом Росстата от 25 февраля 2009 г. № 32 (зарегистрирован в Минюсте России 9 июня 2009 г., рег. № 14062), проводит конкурс на замещение вакантных должностей федеральной государственной гражданской службы в Росстате.

Прием документов осуществляется по адресу: **107450, г. Москва, ул. Мясницкая, д. 39, строение 1.**

Контактный тел.: **632-91-12.**

Начало приема документов для участия в конкурсе будет проводится с 9.00 до 16.45 часов в период с 31 августа по 29 сентября 2009 г.

С подробной информацией о Федеральной службе государственной статистики можно ознакомиться на официальном сайте Росстата: htpp:// www.gks.ru.