Ремонт выборки

А.А.Давыдов, А.О. Крыштановский

Первая публикация статьи: СОЦИС. 1989. № 5. С. 100—105.

Практика проведения социологических исследований показывает, что, как бы тщательно ни был спланирован полевой этап сбора информации, всегда имеют место смещения выборок по социально-демографическим характеристикам, пропущенные ответы в анкетах и некоторые другие моменты, снижающие качество социологической информации.

Для того чтобы свести к минимуму влияние этих нежелательных факторов, в методической литературе [1—3] рекомендуется проводить ремонт выборки. Однако у большинства социологов нет ясного представления о практической реализации этого необходимого этапа социологического исследования. Так, среди 300 исследований, содержащихся во Всесоюзном банке социологических данных ИС АН СССР, лишь в десяти осуществлялся ремонт выборки. Для сравнения отметим, что за рубежом ремонт выборки уже давно стал распространенным методом повышения качества социологической информации.

Причины нашего отставания в применении ремонта выборки очевидны: отсутствие вычислительной техники, специализированного программного обеспечения, методических пособий, недостаточная квалификация исследователей и ряд других факторов. В настоящее время положение меняется в лучшую сторону, поэтому мы считаем разговор о ремонте выборки актуальным.

Что же такое ремонт выборки? В узком смысле — это уравнивание выборочных и генеральных распределений социально-демографических характеристик респондентов. В широком — первичная статистическая обработка данных, включающая коррекцию:

- смещения социально-демографических характеристик респондентов;
- неоднородности массивов данных;
- резко выделяющихся и восстановление пропущенных ответов.

Таким образом, цель ремонта выборки — повышение качества уже собранной информации. Добиться этой цели можно, если использовать избыточную информацию, которая содержится в собранных данных. Это важнейшее положение, к сожалению, редко учитывают социологи.

Рассмотрим общие методологические принципы, на которых базируется логика ремонта выборки. Первый — доминирование неформальных процедур принятия решений. Успех при ремонте выборки достигается благодаря глубокому знанию изучаемой проблемы, процедуры выборочного исследования, ситуации опроса респондентов и т.д., а собственно математические процедуры играют подчиненную роль. Такой подход в корне противоположен мнению, согласно которому ремонт выборки — это недостойные серьезного социолога математические манипуляции, уводящие в сторону от познания социальной реальности.

Второй принцип — оптимизация. Повышение качества собранной информации осуществляется благодаря максимальному уменьшению влияния нежелательных факторов при минимальном искажении, вносимом ремонтом выборки. Поясним это положение примером. Допустим, мы опросили мужчин больше, чем требовалось, и теперь уменьшаем их количество до требуемой квоты. В результате объем выборки может оказаться недостаточным для решения поставленных задач. Значит, сокращать объем можно только до не которого оптимума. При этом надо помнить, что ремонт не заменяет расчета выборки и качественной работы анкетеров. Он лишь облегчает усилия по ее расчету и реализации.

Рассмотрим процедуры ремонта выборки с того момента, когда данные введены в компьютер и «очищены» от ошибок перфорации.

Коррекция неоднородности сбора данных. Неоднородность сбора данных возникает по двум причинам. Во-первых, на полевой стадии исследования практически всегда сбор информации осуществляют несколько анкетеров, различающихся степенью подготовки, социально-демографическими и личностными характеристиками. Кроме того, анкетеры не всегда проводят опрос в одинаковых условиях. Все это оказывает воздействие на ответы респондентов, причем достаточно сильное [4].

Во-вторых, при проведении почтового опроса сбор информации растягивается иногда на несколько недель, и данные, поступившие в разные периоды полевого этапа, могут отражать временные изменения изучаемого явления или разные группы респондентов.

Перед исследователем возникает вопрос: правомерно ли объединять эти данные в один общий массив, если анализировать их вместе нельзя, поскольку в этом случае мы получим существенные искажения? Можно ли рассматривать несколько совокупностей данных в качестве различных выборок, полученных из одной и той же генеральной совокупности [5]. В случае, когда группы данных оказались неоднородными, перед исследователем возникают проблемы коррекции.

Если различия в массивах данных обусловлены влиянием анкетера или условиями опроса, следовало бы провести повторный опрос в данной выборочной совокупности с помощью другого анкетера или анкетеров. Но это в идеале, а на практике — дефицит времени, материальных и людских ресурсов. Кроме того, могут произойти существенные изменения в общественной жизни, и неясно, что мы получим: информацию, которая отражает предыдущий период, или результаты изменений в общественной жизни. Следует так же учитывать, что повторный опрос одних и тех же людей ведет к искажениям, обусловленным психологическими особенностями, а если опрашивать других респондентов со схожими социально-демографическими характеристиками, изменения в оценках могут быть обусловлены новым объектом.

В этой ситуации на помощь может прийти одна из основных процедур ремонта выборки — перевзвешивание данных. Суть ее состоит в следующем: с помощью эмпирически найденных весов так скорректировать данные, чтобы влияние смещений снизить до оптимальных пределов. Величина смещения находится либо с помощью экспертного опроса, либо как отклонение средних значений в подвыборках от среднего значения по всему массиву.

При перевзвешивании данных возникает проблема правильного выбора эталона, по отношению к которому будет осуществляться коррекция неоднородности. Эталон можно выбрать исходя из содержательных соображений либо конструировать его как нечто среднее из тех подвыборок, которые имеются в наличии. Одним из таких эталонов может выступать средневзвешенная величина смещения по всему массиву. В табл. 1 представлены этапы подобной коррекции.

Таблица 1. Коррекция неоднородности данных*

Номер анкетера	**1**	**2**	**3**
Исходное количество анкет	80	54	101
Балл смещения	2	1	3
Средневзвешенная величина смещения	(80 x 2 + 54 x 1 + 101 x 3) : 235 ≈ 2,2		
Процедура расчета веса	2,2 : 2 = 1,1	2,2 : 1 = 2,2	2,2 : 3 ≈0,73
Коррекция количества анкет	80x1,1=88	54x2,2=119	101x0,73=74

* Примечание. Эксперты оценивали смещение с помощью 3-балльной шкалы, где 1 балл — смещение незначительное, а 3 балла — смещение очень значительное.

В результате коррекции объем массива увеличился на 46 анкет (на 20% первоначального объема). Если бы в качестве эталона была выбрана не средневзвешенная величина смещения, как в табл. 1, а минимальное смещение (1 балл), объем скорректированного массива сократился бы на 108 анкет (46% начального объема).

Таким образом, стремление к максимальному уменьшению смещения привело к сокращению исходного массива почти вдвое. В то же время использование средневзвешенной величины смещения в качестве эталона заставило продублировать каждую анкету второго анкетера. И хотя в методической литературе высказывается мнение, что нельзя одну и ту же анкету включать в машинную обработку более 10—11 раз [2], все же дублирование анкет увеличивает влияние индивидуальных особенностей опрашиваемых, которое в каждом конкретном исследовании различно. Поэтому в одном исследовании правомерно увеличить подмассив вдвое, а в другом — нет. Проблема верхних границ дублирования остается открытой, поэтому знание проблемы и здравый смысл — основные критерии принятия правильного решения. Мы рассмотрели коррекцию смещений, вызванных влиянием анкетера.

Аналогично проводится коррекция смещений, обусловленных различием массивов данных во времени.

После того как веса найдены и массив скорректирован', с помощью статистических критериев следует еще раз провести проверку на однородность. В случае, если подмассивы снова оказались неоднородными, требуется новое перевзвешивание, но уже с другими весами, и затем проверку надо повторить. Если скорректированные подмассивы опять окажутся неоднородными, их следует обрабатывать отдельно.

Коррекция распределений социально-демографических характеристик респондентов. После сбора информации практически всегда наблюдается смещение социально-демографических характеристик опрошенных, по сравнению с генеральной совокупностью. Прежде чем приступать к коррекции, полезно выявить влияние социально-демографических признаков на ответы респондентов. Этот анализ может быть осуществлен с помощью двумерных таблиц сопряженности или множественного номинального анализа.

Например, нами установлено, что социально-демографические признаки слабо связаны с ответами об удовлетворенности работой и жизнью, оценкой темпов перестройки, одобрением деятельности политических лидеров, оценкой внешнеполитических событий и др. Для этих индикаторов перевзвешивание по социально-демографическим характеристикам не нужно.

Возможны три ситуации. Первая — ответы респондентов не связаны с социально-демографическими характеристиками, в этом случае коррекция не проводится. Вторая ситуация — какая-то социально-демографическая характеристика, например пол, тесно связана со всеми содержательными вопросами, или третья — разные вопросы могут быть связаны с различными характеристиками. В этом случае коррекция проводится по схеме, описанной в [3].

Из табл. 2 следует, что в результате коррекции количество мужчин уменьшилось на 28 человек, и на столько же увеличилось число женщин. Дублирование анкет женщин основывается на базовом принципе выборочного метода [б], согласно которому каждый индивид несет всю информацию, представленную в его социально-демографической группе. В табл. 2 приведен пример, когда на ответы респондентов оказывает влияние только одна характеристика — пол. Практика показывает, что так бывает далеко не всегда. Значительно чаще встречается ситуация, когда на ответы респондентов оказывают влияние две, например возраст и образование, или три и более социально-демографические характеристики. Для этого случая одним из авторов статьи разработаны метод и соответствующие программы для ЭВМ, рассчитывающие веса для нескольких

признаков одновременно [7]. Здесь отметим следующее: использование данных программ в ИС АН СССР показало их высокую эксплуатационную надежность, а главное — простоту в обращении.

Таблица 2. Коррекция по одной социально-демографической характеристике.

Пол	Выборочная совокупность		Генеральная совокупность, %	Расчёт веса	Скорректированная совокупность	
	численность	%			численность	%
Мужчины	100	66,6	48	48 : 66,6=0,72	100x0,72=72	48
Женщины	50	33,4	52	52 : 33,4-1,56	50x1,56=78	52

После коррекции выборочных распределений социально-демографических признаков можно приступать к следующему этапу ремонта выборки — коррекции резко выделяющихся и восстановлению пропущенных ответов.

Коррекция резко выделяющихся ответов респондентов. В практике опросов общественного мнения встречаются ответы респондентов, которые сильно отличаются от основной массы ответов. Это может быть обусловлено ошибкой самого респондента или ошибкой регистрации ответа интервьюером, иногда—особенным мнением респондента или резким изменением условий опроса. Установить истинную причину отклонения практически невозможно. Резко выделяющиеся ответы затрудняют анализ данных поэтому вполне естественно стремление их как-то найти и скорректировать

Выявлению резко выделяющихся наблюдений посвящено большое количество научных публикаций (например, [5]), поэтому мы не будем подробно останавливаться на этой задаче, и рассмотрим проблему коррекции подобных наблюдений. Самый простой способ — удалить данный ответ или всю анкету и дальнейшего анализа. Эта возможность предусмотрена в пакетах «Социолог» BMDP-79, SPSS и ряде других. Однако когда объем выборки невелик, это обходится слишком дорого, особенно если резко выделяющихся ответов много.

Второй способ — отнесение резко выделяющихся ответов к градации «другое». Этот прием применяется при кодировке открытых вопросов и с успехом может быть использован при коррекции резко выделяющихся ответов, поскольку «отнесение» таких ответов в одну градацию обеспечивает наполнение и делает возможным дальнейший анализ.

Третий способ—уменьшение дробности шкалы. Например, Г.И. Саганенко отмечает, что шкалу в пять-семь-девять градаций почти всегда приходится сводить к трем-четырем [б]. Эту задачу можно решить с помощью» статистических критериев, например критерия Фишера, который показывает, значимо ли различаются доли ответов респондентов. Наш опыт свидетельствует, что уменьшение дробности шкалы позволяет эффективно бороться с резко выделяющимися ответами.

Коррекция пропущенных ответов. Данный вид смещений возникает чаще всего в открытых вопросах и вопросах табличного типа. Самый простой способ коррекции — исключение из дальнейшего анализа пропущенных» ответов или всей анкеты. Если объем выборки большой, это весьма рациональный подход. В условиях выборок малых и средних объемов распространенными способами коррекции являются: отнесение пропущенного ответа к градации «затрудняюсь ответить», замена пропущенного ответа каким-либо средним значением, рассчитанным по имеющимся данным, или значением, вычисленным с помощью регрессии. Названные процедуры реализованы в пакетах «Социолог», BMDP-79, SPSS и ряде других. Выбор того или иного способа коррекции пропущенных ответов в значительной мере зависит от последующего анализа данных. Например, при расчете одномерного частотного распределения пропущенный ответ логично отнести к категории «затрудняюсь ответить». Однако если предполагается

факторный анализ, такой подход неприемлем, поскольку эта категория исключается из обработки.

При планировании факторного анализа более естественно заполнение пропусков модальным, медианным или среднеарифметическим значением, вычисленным по всему массиву или в социально-демографической группе того респондента, который не ответил на вопрос. Предполагается, что мода, медиана или среднеарифметическое значение отражают общую тенденцию, а другие ответы, отклоняющиеся от этих значений, обусловлены влиянием личностных особенностей респондентов, различиями в ситуации опроса и другими случайными факторами.

При планировании логлинейного анализа коррекция пропущенных ответов осуществляется другим способом. Напомним, что в логлинейном анализе от частоты в каждой ячейке таблицы сопряженности берется логарифм, и, если там нет наблюдений, то, строго говоря, данный анализ невозможен. Поэтому в статистической литературе рекомендуют перед проведением логлинейного анализа в каждую ячейку таблицы сопряженности добавить некоторое небольшое число, как правило, в интервале от 0,25 до 1,00 [8], что позволяет вычислить логарифм при отсутствии ответа. Доказано, что подобная «добавка» не сказывается на качестве результата [9]. (Данная процедура реализована в программе логлинейного анализа пакета BMDP-79, где в каждую ячейку таблицы сопряженности добавляется число 0,5.)

До сих пор мы рассматривали ситуации, когда пропущен содержательный ответ. А что делать, если отсутствует какая-либо социально-демографическая характеристика? В этом случае можно поступить так: если социально-демографические характеристики не связаны с содержательными ответами, то анкете с пропущенными значениями следует присвоить наиболее часто встречающиеся в выборке социально-демографические характеристики, либо определить их случайным образом или пропорционально (если таких анкет много). Если же связь есть, следует определить, к ответам какой группы (например, мужчин или женщин) ближе ответы в анкете, где графа «пол» не указана, и внести этот признак.

Итак, мы осуществили ремонт выборки, и теперь следует оценить смещения, вносимые самим ремонтом. Для этой цели нужно найти эталон, по отношению к которому будет рассчитываться смещение. Возможны два эталона— внутренний и внешний. Процедура построения внутреннего эталона может быть следующей: из выборочной совокупности формируется небольшая подвыборка, в которой практически отсутствуют смещения по социально-демографическим признакам, резко выделяющиеся и пропущенные ответы. По данной подвыборке рассчитываются процентные распределения, связи и т.д., а затем сравниваются с данными, полученными после ремонта.

Мера отклонения от результатов эталонной подвыборки и будет выступать показателем смещения, вносимого ремонтом. Допустимость смещений легко выявляется с помощью статистических показателей, например, %2 распределения. Процедура построения внешнего эталона несколько иная. Во время полевого этапа данные собираются по двум выборочным планам. Один — основной, по которому будет осуществляться ремонт выборки и дальнейший анализ, а второй — для создания эталона. Подразумевается, что эталонная подвыборка не участвует в общем, анализе, собирается особенно тщательно, а распределения социально-демографических характеристик точно соответствуют распределению в генеральной совокупности.

С нашей точки зрения, предпочтение следует отдавать внешнему эталону, поскольку при построении внутреннего могут возникать сложности, обусловленные ограниченным объемом выборки [б].

В заключение отметим еще одно принципиальное положение. Если данных много, ремонт выборки может осуществляться за счет сокращения выборочной совокупности. Это наиболее рациональный подход к ремонту выборки, поскольку данная стратегия не

опирается ни на какие дополнительные допущения. Если объем выборки незначителен, для ее ремонта нужно принимать ряд дополнительных допущений, которые не следуют из собранного материала и истинность которых трудно проверить. Таким образом, возникает дилемма: опрашивать большое количество респондентов, не ; беспокоясь о качестве, в надежде на «капитальный» ремонт выборки, или опрашивать значительно меньшее количество респондентов, но с высоким качеством, предполагая «косметический» ремонт. Ответ следует искать в размере затрат (материальных, временных и др.), вытекающих из каждого решения, в особенностях изучаемой проблемы, целях и задачах исследования и ряде других факторов.

Литература

1. Джессен Р. Методы статистических обследований. М.: Финансы и статистика, 1985.
2. Петренко Е.С., Ярошенко Т.М. Социально-демографические показатели в социологических исследованиях. М.: Статистика, 1979.
3. Процесс обработки данных анкетных опросов на ЭВМ. М.: ИС АН СССР, 1985.
4. Погосян ГА. Метод интервью и достоверность социологической информации. Ереван: Изд-во АН Армянской СССР, 1985.
5. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.
6. Саганенко Г.И. Надежность результатов социологического исследования. Л.: Наука, 1983.
7. Крыштановский А.О., Кузнецов А.Г. Перевзвешивание выборки // Комплексный подход к анализу данных в социологии. М.:ИС АН СССР, 1988.
8. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Т. 1. М.: Финансы и статистика, 1982.
9. Аптон Г. Анализ таблиц сопряженности. М.: финансы и статистика, 1983.