Ю.Н.Толстова,
И.К. Зангиева

Национальный исследовательский университет Высшая школа экономики

ПОНЯТИЕ СЛУЧАЙНОСТИ И ПРОБЛЕМА ПРОПУСКОВ ДАННЫХ В СОЦИОЛОГИИ

1.Введение

В настоящей статье мы возвращаемся к теме вероятностного порождения исходных данных в социологии, рассмотрение которой было начато нами ранее [1]. Соответствующие положения будут рассмотрены в преломлении к проблеме, очень важной самой по себе – проблеме «борьбы» с пропущенными данными. Имеющиеся в литературе наработки по этой проблеме носят довольно абстрактный характер. Опыт показывает, что при их использовании в социологии возникает масса вопросов, часть из которых будет рассмотрена в данной статье. Актуальность такого рассмотрения подтверждается также и тем, что в отечественной социологической литературе соответствующие вопросы не рассматриваются (в русско-язычной литературе те достижения западных авторов, о которых мы собираемся говорить, не описываются)[1]. Кроме того, мы попытаемся более четко, чем это делается в литературе, обрисовать логику процесса ликвидации пропусков, соединив её с логикой статистической парадигмы. Это, на наш взгляд, позволит избежать некоторых недоразумений, нередко возникающих при внедрении методов «борьбы» с пропусками в практику социологических исследований. Будет также показано, что некоторые рассматриваемые положения тесно сопрягаются с отдельными важными методическими аспектами социологического исследования.

2. Подходы к «борьбе» с пропусками. Необходимость использования статистической парадигмы (предположений о вероятностном порождении исходных данных) для их реализации

[1] Наш опыт говорит о том, что даже в тех редких случаях, когда наши социологи пытаются использовать западные наработки, они очень часто делают ошибки именно из-за отсутствия достаточной глубоких знаний по оценке роли случайности того или иного факта в процессе анализа данных и в понимании смысла случайных и неслучайных событий.

Известно, что проблема наличия пропусков в данных[2] очень остра для социолога. Наличие хотя бы одного пропуска препятствует использованию почти всех известных методов анализа данных. Исключение составляют только расчет одномерных выборочных распределений и мер средней тенденции, которые могут быть рассчитаны с учетом либо всех элементов выборки, либо т.н. актеров, т.е. тех, у которых нет пропуска в значении рассматриваемого признака.

В литературе [2, 3, 4 Platek 1980; Sande I. 1982; Rubin 1987] предлагается три способа «борьбы» с пропусками[3]: исключение документов с пропусками, взвешивание выборки с целью замены пропущенных данных искусственно повторенными полными наблюдениями и искусственное заполнение отдельных пропусков. Первый вариант вряд ли может устроить любого исследователя: выбрасывание значительной части собранной информации (а в социологии пропусков бывает много), естественно, может привести к неадекватности реальности выводов, полученных на основе анализа оставшихся данных.

Рассмотрим два другие способа ликвидации пропусков. Оба заставляют нас использовать статистический подход к получению нового знания. Представляется очевидным, что, говоря о том или ином способе «восстановления» пропущенных данных, естественно переводить все встающие здесь вопросы на статистический язык. Вряд ли в принципе можно говорить о точном восстановлении пропущенного значения какого-либо признака для конкретного респондента. Как-то восстановить пропущенную информацию можно только на уровне распределения: можно стремиться к тому, чтобы так или иначе восстановленные данные имели то же распределение, что и изначально полные. Это зачастую оказывается вполне возможным, что по существу

[2] В социологии пропуск в данных чаще всего означает отсутствие ответа респондента на тот или иной заданный ему вопрос анкеты, и именно такие ситуации мы будем использовать в качестве примеров; набор данных, отвечающих одному респонденту, называем наблюдением; используем понятия полного и не полного наблюдения в соответствии с тем, ответил ли соответствующий респондент на все предлагаемые ему вопросы или нет.

[3] Те методы ликвидации пропусков, о которых мы будем говорить, иногда называют методами, употребляющимися после сбора данных, и противопоставляют им методам, тоже направленным на ликвидацию пропусков, но употребляющимся до сбора данных. Последняя категория методов – это методы выработки таких методических подходов (составления анкеты, поведения интервьюера и т.д.), которые стимулируют респондента дать ответ. Мы эти методы затронем лишь в малой степени.

используют известные методы «борьбы» с пропусками.. Рассмотрим, ак в нашем случае будет выглядеть статистический подход.

В работе [1] мы обсуждали сущностные и терминологические опросы, связанные с принятием статистической парадигмы, родемонстрировали наличие большого количества неясностей в бласти установления соответствия между математическим ормализмом и содержательными представлениями социолога об зучаемых явлениях и пришли к выводу, что использование татистического подхода имеет смысл считать тождественным ринятию вероятностной модели порождения исходных данных. Мы не удем повторять сказанное в упомянутой статье, укажем лишь главные оменты, которые объединим с рассуждениями о понятии случайности ропусков в работах, посвященных рассмотрению проблемы ропущенных данных.

Вероятностная модель порождения исходных данных чаще всего пирается на использование следующих предположений[4]. Понятие енеральной совокупности осмыслено. Оно отождествляется с аспределением изучаемой многомерной случайной величины (мы удем говорить о совокупности распределений набора одномерных лучайных величин). Из генеральной совокупности возможно ерпать» бесконечное количество случайных выборок и наша выборка одна из них; каждый рассматриваемый признак – это выборочная еализация случайной величины (отождествляемой с неким аспределением вероятностей встречаемости её значений), все аблюдаемые, частоты это выборочные оценки генеральных ероятностей. Под ***элементарными событиями*** в такой ситуации онимаем события, состоящие в том, что такой-то признак принимает акое-то значение. Другими словами, случайное событие – это ыборочное значение одной из случайных величин, определяющих ашу генеральную совокупность. Какие же случайные величины ризнаки мы должны принять во внимание? Ясно, что, в первую ередь, те, которым отвечают вопросы нашей анкеты, но не только: не адо забывать о наших пропущенных данных. Введем ряд определений, оторые соответствуют известной из литературы логике процесса ивидации пропусков, но, насколько нам известно, нигде в четком виде е сформулированы.

С формальной точки зрения мы, говоря о вероятности, опираемся на звестное колмогоровское определение вероятностного пространства – ожады, состоящей из множества элементарных событий, алгебры событий и ероятности, т.е. меры, заданной на событиях. Но фактически используем одно з самых популярных практических воплощений этой теоретической модели.

Будем полагать, что в анкете задействованы две группы случайных величин, два вида распределений, каждое из которых отвечает одному анкетному вопросу. Случайные величины первой группы отражают вероятности встречаемости всех значений признака. Назовем такие ***случайные величины величинами первого типа***. Случайные величины второй группы отражают вероятности ответа и неответа на каждый вопрос, такие случайные величины являются дихотомическими, будем говорить о них как о случайных величинах вида «ответ-неответ». Назовем такие случайные величины ***величинами второго типа.*** Их введение означает распространение статистического подхода на ту часть работы социолога, которая связана с пропущенными данными. Как мы покажем ниже, изучение такого рода случайных величин и дает основание грамотного выбора способа «борьбы» с пропусками, определения того, можем ли мы взвешивать части выборки, искусственно заполнять пропуски, либо же нам ничего не остается, как только выбросить наблюдения с пропущенными значениями каких-либо признаков.

3. Об относительности понятия случайности. Внешние и внутренние факторы, влияющие на вероятность неответа

Рассматриваемые методы работы с пропусками существенным образом опираются на своеобразное понятие случайности пропусков [5]. Наш опыт показывает, что это понятие далеко не всегда адекватно воспринимается социологами. Уточним его путем сведения к наличию случайных величин второго типа. Но для того, чтобы это сделать, необходимо сказать пару слов о самом понятии случайности и выделить несколько методических моментов, связанных с этим понятием.

Во-первых, понятие случайности носит относительный характер. Одно и то же явление может быть случайным относительно каких-то одних факторов и неслучайным – относительно других. Известно классическое выражение Ф.Энгельса о том, что «всякая случайность есть пересечение необходимостей». Сведение случайности к необходимости осуществлялось в классической (восходящей к Ньютону) науке. Ученые полагали, что если мы все будем знать про анализируемый процесс, то случайностей не будет и, стало быть, не будет потребности прибегать к понятию вероятности. Приведем классический пример с подбрасыванием монеты. Представим себе, что монета падает. Если бы исследователь смог учесть все факторы, определяющие её движение – малейшие колебания воздуха и температуры, микроскопические движения пальцев, бросающих монету

и т.д. – то он точно знал бы, на какую сторону монета упадет. Понятие вероятности не нужно, поскольку потребность в его использовании возникает из-за недораскрытости законов природы. Но в наше время стало ясно, что случайность не может быть убрана из науки. Причин, определяющих любое явление, бесконечно много (да и до сих пор неясно, что такое причина). Каждое явление описывается бесконечным количеством признаков и уже хотя бы поэтому не может быть полностью описано с помощью связывающих эти признаки закономерностей (коих тоже – бесконечное число). Мы можем изучать природу всё глубже, выявлять все более глубокие закономерности, но все те обстоятельства, которые детерминируют даже самое простое явление, нами не будут изучены никогда. Поэтому случайность всегда будет присутствовать в науке. А где случайность – там и вероятность, и теория вероятностей, и математическая статистика, и статистический подход к получению нового знания.

Во-вторых, мы должны четко понимать, в рамках какой модели изучаемого явления работаем. Всегда необходимо давать себе отчет в имеющихся ограничениях. В наших рассуждениях моделью респондента является рассматриваемая анкета. Используя её, мы как бы считаем, что всё, что в поведении респондента нас интересует (например, причины, определяющие его ответ, либо неответ на какой-либо анкетный вопрос), целиком зависит (не зависит) от рассматриваемых качеств респондента, т.е. от его ответов на этот и другие вопросы той же анкеты. Подчеркнем, что мы сейчас готовим почву для разговора о способах «борьбы» с пропусками. Как мы увидим, смысл случайности пропусков определяется именно характером реакции респондента на предъявляемые ему анкетные вопросы. Это все время надо помнить. Но, естественно, модель всегда может быть расширена. Ниже мы осуществим такое расширение, перейдя к рассмотрению некоторых причин пропусков, не связанных с ответами на вопросы анкеты, а выходящих за пределы того, что в ней отражено, связанных, скажем, с методической организацией исследования, с психологическими характеристиками респондента, с социальными условиями, в которых он живет и т.д. Для удобства назовем причины пропусков, связанные с «анкетными» свойствами респондента, ***внутренними факторами***. А причины неответов, не связанные с такими свойствами – ***внешними факторами***. В социологической литературе имеется довольно много работ, посвященных выявлению и анализу внешних факторов [6-10].

Подчеркнем, что в число внешних факторов попадают не только такие признаки, которые вообще не имеют никакого отношения к анкете (скажем, психологические характеристики респондента), но и,

скажем, ошибки в формировании списка предлагаемых ответов на некий вопрос анкеты, если эти ошибки исходят от самого исследователя и не зависят ни от каких случайных величин как первого, так и второго типа. Скажем, что если рассматриваемый вопрос сформулирован некорректно и за счет этого на него многие респонденты не отвечают, то эта некорректность относится к числу внешних факторов, поскольку здесь вероятность неответа не связана с «анкетными свойствами респондента.[5]

4. Уровни случайности пропусков в теории пропущенных данных и их связь со случайными величинами второго типа.

В качестве фактора, определяющего выбор допустимого способа ликвидации пропусков, была выдвинута степень случайности пропуска: возможность рассматривать пропуск неслучайным, случайным и полностью случайным [5]. Рассмотрим это более подробно, связав указанные виды случайности с традиционными представлениями о случайных событиях, являющихся реализациями значений некоторых случайных величин, а понятие случайности неответа на некий вопрос – с осмысленностью соответствующей случайной величины второго рода.

Событие, состоящее в ответе (неответе) респондента на некий вопрос анкеты, является выборочной реализацией соответствующей этому вопросу случайной величины второго типа. И упомянутое событие можно будет считать ***случайным***, если эта случайная величина действительно будет случайной величиной, т.е. если для неё будет определено вероятностное распределение, будут известны вероятности ответа и неответа на рассматриваемый вопрос. Очевидно, это будет иметь место, когда мы сможем интерпретировать частоты ответов (неответов) как выборочные оценки вероятностей значений нашей дихотомической случайной величины, т.е. вероятностей ответа (неответа) на рассматриваемый вопрос. То же можно выразить и так: случайными будем называть пропуски в значениях конкретного признака при условии, что может быть определена вероятность ответа или неответа респондентов на соответствующий вопрос анкеты. Другими словами, пропуски значений рассматриваемого признака могут

---

[5] Конечно, в нашем примере может оказаться, что некорректность вопроса заставляет не отвечать на него лишь какие-то определенные категории респондентов, заданные анкетой: скажем, на «ответ-неответ» мужчин эта некорректность не влияет, а вот женщины не хотят отвечать на этот вопрос. Учет подобных фактов говорит о необходимости определенных теоретических разработок, на которые мы здесь не претендуем.

считаться случайными, если соответствующий дихотомический признак «ответ-неответ» - это случайная величина.

Казалось бы, проблемы здесь нет: посчитал долю неответивших на рассматриваемый вопрос – вот и оценка вероятности неответа. Но это кажущаяся простота. В действительности здесь появляется еще одна проблема – проблема однородности изучаемой совокупности респондентов: указанные доли могут сильно различаться для разных категорий респондентов. Как мы увидим ниже, именно это обстоятельство привело к рассмотрению разных уровней случайности. К обсуждению проблемы однородности вернемся в п.8.

Пропуски, возникающие при ответе на какой-либо выбранный анкетный вопрос, называются ***полностью случайными*** (missing completely at random — MCAR), если вероятность их возникновения не зависит ни от каких внутренних факторов: ни от истинного значения данного признака (респондент может не отвечать, но мы предполагаем, что какое-то значение соответствующего признака, пусть не известное нам, он имеет), ни от значений других признаков. «Полная» случайность пропусков дает нам основание считать, что случайная величина второго типа для рассматриваемого признака определена, т.е. ее распределение осмыслено и может быть оценено с помощью выборочной доли неответивших на рассматриваемый вопрос.

Пропуски называются ***случайными*** (missing at random — MAR, будем говорить о таких пропусках как о ***«просто случайных»***), если вероятность их возникновения обусловлена известными значениями других переменных, но не связана с переменной, значение которой пропущено. Например, вероятность неответа на вопрос о возрасте может быть разной для мужчин и для женщин. В таком случае случайная величина второго типа для рассматриваемого признака («ответ-неответ») не будет осмыслена для всей рассматриваемой совокупности респондентов, соответствующее распределение не является единым для неё и поэтому пропуски мы не можем считать случайными,. Но, стоит только разделить совокупность на подсовокупности, отвечающие тем качествам респондентов, от которых зависит указанное распределение, то положение исправится. Так, в приведенном примере своя случайная величина второго типа будет существовать для мужчин и женщин в отдельности. Другими словами, первоначальная совокупность респондентов оказалась неоднородной, в ней по сути скрывались две выборки, каждая из которых соответствовала своей генеральной совокупности. Первоначальная неслучайность оказалась относительной, при разделе совокупности на две превратилась в случайность.

Пропуски при ответе на какой-либо вопрос ***неслучайны*** (not missing at random — NMAR), если закон распределения соответствующей случайной величины второго типа нельзя определить из-за того, что вероятность ответа (неответа) на рассматриваемый вопрос зависит от того, какое значение рассматриваемого признака в действительности имеет респондент (скажем, возможна ситуация, когда пожилые люди с большей вероятностью не дают ответа, чем молодые, на вопрос о возрасте). Будем считать, что случайная величина второго типа для рассматриваемого анкетного вопроса не существует. Ниже (см. п. 8) мы покажем, однако, что и здесь понятие случайности относительно: при определенных предположениях неслучайные пропуски могут стать случайными.

Для ясности заметим, что при таких определениях «полная случайность» пропусков ответа на какой-либо фиксированный вопрос в анкете означает независимость распределения соответствующей случайной величины второго типа («ответ-неответ») от внутренних факторов, опирающихся на величины первого типа.[6]; «простая случайность» ответов на какой-либо вопрос означает зависимость соответствующей этому вопросу случайной величины второго типа от распределений первого типа, отвечающих другим вопросам анкеты. Другими словами, вместо случайной величины типа «ответ-нетвет» для рассматриваемого вопроса мы должны иметь дело с условными распределениям этой величины, причем условие должно задаваться значениями величин первого типа для других вопросов анкеты. Неслучайность пропусков в ответах на какой-то вопрос означает зависимость случайной величины второго типа для этого вопроса от случайной величины первого типа для него же.

Все сказанное выше, помимо всего прочего, иллюстрирует относительность понятия случайности: неслучайные, на первый взгляд, пропуски в одних ситуациях легко превратить в случайные (см. определение «просто случайности» ниже), а в других – это сделать невозможно (см. понятие неслучайности пропусков ниже).

Описанное выделение видов случайностей является практически полезным: оно как бы стимулирует исследователя в явном виде выделять те рассмотренные в [1] ситуации, когда случайная величина фактически не существует. Это иллюстрируется сказанным выше: наша дихотомическая случайная величина второго типа может перестать

---

[6] Здесь хотелось бы ввести понятие внутренних факторов, опирающихся на случайные величины второго типа (т.е. отразить то, что вероятность ответа на какой-либо один вопрос может зависеть от вероятности ответа на другой вопрос), но это в литературе мы не встречали.

быть случайной величиной (т.е. соответствующее распределение вероятностей может отсутствовать) по двум причинам: или за счет того, что фактически она распадается на две случайные величины, каждая из которых отвечает своему набору значений других признаков анкеты (к примеру, рассматриваемое распределение для мужчин имеет один вид, а для женщин – другой; это отвечает «просто случайности» пропусков) или за счет того, что соответствующая вероятность неответа на рассматриваемый вопрос не может быть вообще определена, она не существует, поскольку её значение зависит от разных вариантов возможных ответов (это отвечает ситуации неслучайности пропусков).

В заключение параграфа свяжем понятие случайности пропусков с понятием генеральной совокупности. «Полная» случайность эквивалентна тому, что мы имеем дело с единой генеральной совокупностью. А это, в свою очередь, выражается в том, что рассматриваемые случайные величины (и первого, и второго типа) имеют для всей совокупности (в частности, и для полных наблюдений, и для наблюдений с пропусками) единое распределение. «Простая» случайность значит единство соответствующих условных распределений. Неслучайность означает отсутствие генеральной совокупности, поскольку на всем множестве респондентов явно не наблюдается единство распределений для случайных величин второго типа, а выделить совокупности с одинаковыми распределениями мы не можем.

5. «Ликвидация» пропусков путем взвешивания

На примере покажем, что делает взвешивание.

Сначала рассмотрим ситуацию, когда пропуски – *полностью случайные*. Предположим, что речь идет об отсутствии ответа на вопрос о возрасте. Для простоты пока предположим, что на все остальные вопросы все респонденты ответили. В таком случае, как мы уже говорили, определено распределение случайной величины второго типа («ответ-неответ») применительно к возрасту, и можно полагать, что процент не ответивших является хорошей выборочной оценкой генеральной вероятности неответа на вопрос о возрасте. Скажем, оказалось, что 10% респондентов, вошедших в выборку, не указали свой возраст. Считаем, что вероятность неответа (подчеркнем, не зависящая ни от каких других отраженных в анкете качеств респондента, одинаковая для всех) на соответствующий вопрос анкеты равна 0,1. Логика дальнейших действий такова. Предполагаем, что

распределение тех респондентов, которые скрыли свой возраст, по всем рассматриваемым признакам такое же, как и у тех, кто сообщил о своем возрасте (у нас нет оснований сомневаться в этом). Последних – 90% от нашей первоначальной выборки. Искусственным образом расширяем их совокупность: случайным образом выбираем 11% среди ответивших о возрасте (это будет примерно то же количество респондентов, что и 10% от первоначальной выборки) и каждого из выбранным респондентов «удваиваем». Получившиеся двойники послужат как бы заменой тех респондентов, у которых возраст был пропущен. К массиву людей, состоящих из тех, у кого изначально не было пропусков с добавленными к ним 11% двойников применяем те методы анализа данных, которые нас интересуют и считаем, что мы получаем те выводы, которые получили бы, если бы неответы на вопрос о возрасте отсутствовали.

Совершенно та же логика используется, если у нас имеются пропуски по нескольким вопросам. Не будем соответствующими рассуждениями загромождать изложение. Подчеркнем, что при реализации рассматриваемой техники мы как бы восстанавливаем не сами пропущенные данные, а те распределения, которые ассоциируются с рассматриваемой генеральной совокупностью.

Теперь рассмотрим *просто случайные пропуски*. Предположим, например, что вероятность неответа респондента на вопрос о возрасте зависит от пола: среди женщин 20% склонны не давать ответ о своем возрасте, а среди мужчин – 10%. Положим также для определенности, что в нашей совокупности 30% мужчин и 70% женщин, а общее количество респондентов равно 1000. Сказанное отражено в табл. 1.

Таблица 1. Пример распределения вероятности неответа, при котором пропуски могут быть просто случайными

	Мужчины (кол-во, %)	**Женщины (кол-во, %)**	**Всего (кол-во, %)**
Всего в выборке	300 (100)	700 (100)	1000 (100)
Не ответили на вопрос о возрасте	30 (10)	140 (20)	179 (17,9)

Здесь уже возникают сомнения в том, что распределения случайных величин первого типа одинаковы у мужчин и женщин (женщины имеют большую склонность скрывать свой возраст, чем

мужчины! И причины этого вполне могут отражаться в их ответах на другие вопросы анкеты). Становится бессмысленной постановка вопроса о сходстве распределений у всех ответивших и всех не ответивших на вопрос о возрасте. И таким же бессмысленным представляется заполнение 17,9 % неответивших (179 человек) произвольными случайно выбранными респондентами из генеральной совокупности. Но сомневаться в том, что распределения одинаковы у ответивших и неответивших мужчин и ответивших и не ответивших женщин у нас нет оснований. И вполне логичной выглядит реализация описанной ранее процедуры «восстановления» пропусков путем замены наблюдений с пропущенными данными другими, полными наблюдениями, отдельно для мужчин и женщин. Имеются основания полагать, что взвешивая соответствующим образом отдельно представителей каждого пола140 полных наблюдений женщин и 30 наблюдений мужчин, мы сохраним вид всех генеральных условных распределений (условие – фиксация пола).

В случае *неслучайных пропусков* мы не можем заменять неполные наблюдения полными. Применительно к нашему примеру, неслучайность пропуска возраста означает, что, скажем, старые люди с большей вероятностью не раскроют свой возраст, чем более молодые. В таком случае распределение соответствующей дихотомической переменной «ответ-неответ» не определено! Стало быть, мы не можем считать эту переменную случайной величиной второго типа. Использовать такую фикцию для замены данных с пропусками полными данными вряд ли возможно. Поясним это с помощью таблицы 2.

Таблица 2. Пример вопроса, для которого не существует единого распределения вероятности неответа

Возраст	**15-30**	**30-45**	**45-60**	**>60**
Вероятность неответа на вопрос о возрасте	5 %	10%	20%	30%

Допустим, что на вопрос о возрасте всего не ответили 20% респондентов. Если мы заместим массив соответствующих неполных наблюдений случайным образом отобранными полными, мы почти наверняка исказим фактическое распределение по возрасту

респондентов, не ответивших на этот вопрос по неизвестному нам возрасту. Ведь, если пожилые люди в большей мере не отвечали, что среди неответивших, вероятно, будет больше пожилых, чем в остальной выборке.

Приведем еще один пример того, что восстановление распределений при неслучайных пропусках невозможно. Рассмотрим признак «доход». Предположим, что доход скрывают самые богатые. В числе же наших полных наблюдений остались в основном люди со средним и низким доходом. И как бы мы ни взвешивали полные наблюдения, какой бы замечательный метод для этого ни выбрали, мы можем лишь подменить доходы богатых доходами более бедных слоев общества. Более того, восстановить распределение мы не сможет даже в том случае, если будем пользоваться другим способом «ликвидации» пропусков – искусственным их заполнение (о чем пойдет речь ниже). Другого не дано. Другой информации, кроме информации о средних и малых доходах, у нас просто нет. Естественно, результатом такого «заполнения» у нас будет информация, уводящая нас в сторону от реальной генеральной совокупности.

6. «Ликвидация» пропусков путем их искусственного заполнения

Известно несколько десятков алгоритмов заполнения пропусков. В ряде работ обсуждаются некоторые условия их применимости, предлагаются и определенные подходы к сравнению алгоритмов (в частности, относительно эффективности этих алгоритмов с точки зрения устойчивости результатов анализа тех данных, которые получились в результате заполнения пропусков с помощью сравниваемых алгоритмов). Однако в этой области существует много нерешенных вопросов, в том числе касающихся привязки известных наработок к специфике социологических задач. В настоящей статье мы рассмотрим только один аспект процесса заполнения пропусков: на примере одного из самых популярных методов заполнения - регрессионного моделирования пропусков [11, 12] - покажем, что в основе соответствующих алгоритмов лежит предположение о том, что мы работаем с единой генеральной совокупностью, что выражается, во-первых, в предположении об осмысленности распределений всех рассматриваемых случайных величин для всей совокупности. Подчеркнем, что здесь мы не делаем различия между «полной» и «простой» случайностью. То, что распределение, скажем, случайной величины второго типа для возраста может быть разным для мужчин и для женщин, не создает исключения, поскольку здесь речь идет об условных распределения, тоже осмысленных для всей совокупности –

одно и тоже распределение имеет место для всех респондентов, удовлетворяющих условию «быть женщиной», и то же для условия «быть мужчиной». (другими словами, мы предполагаем, что все наши распределения одинаковы внутри групп, выделенных по полу).

Предположим, что респондент не дал ответа на вопрос «Насколько Вы удовлетворены своей работой?» с семью вариантами ответов от «Совершенно не удовлетворен» до «Полностью удовлетворен» с приписыванием респонденту, соответственно, чисел от 1 до 7. Предположим, что мы может считать эту шкалу по крайней мере интервальной[7]. Предположим также, что на основе полных данных мы нашли хорошую регрессионную зависимость $y = f(x_1, x_2, \dots, x_n)$, где y – это удовлетворенность работой, а независимые переменные – это, скажем, уровень творческой составляющей работы, возможность для респондента раскрыть на работе свою личность, стаж, удовлетворенность товарищами по работе, начальством и т.д. Предположение о единстве системы наших случайных величин для полных и неполных данных заставляет думать, что та же зависимость будет иметь место и для данных с пропусками ответов на вопрос об удовлетворенности (напомним, что регрессионная зависимость, равно как и любая другая закономерность, найденная с помощью известных методов анализа данных, полностью определяется совокупностью рассматриваемых случайных величин, отождествляемых нами с генеральной совокупностью). И, если для респондента с пропуском известны все наши независимые переменные, мы подставляем их в уравнение регрессии и находим значение зависимой переменной, которое и приписываем респонденту с пропуском, тем самым «латаем дырку». Помимо опоры на предположение о принадлежности всех наших данных одной и той же генеральной совокупности, мы опираемся на предположение о том, что, поставив результат регрессионного анализа на место пропущенной информации, мы не сильно исказили истинные (не известные нам) данные и с большой вероятностью сохранили «истинное» распределение. Ведь мы поставили вместо пропущенной удовлетворенности значение, являющееся средним для того набора значений независимых переменных, который отвечает рассматриваемому респонденту с пропуском. Если соответствующее условное распределение у-ков является нормальным (что предполагается в классических регрессионных моделях), то мы подставили самое вероятное значение зависимой переменной при данной наборе значений зависимых.

[7] Обоснование такой возможности можно найти в работе [13]

Примерно такие же рассуждения справедливы и для других алгоритмов заполнения пропусков.

Рассмотрим теперь ситуацию с неслучайными пропусками. Вернемся к примеру с богатыми и бедными респондентами, о котором шла речь в конце п. 5. Как мы там отмечали, в рассматриваемой ситуации имеются все основания полагать, что распределение по доходу у ответивших на соответствующий вопрос в анкете и не ответивших на него, будет различным: среди ответивших доля богатых будет меньше, чем среди неответивших. Алгоритмы заполнения пропусков будут использовать имеющиеся данные, т.е. данные, относящиеся в основном к людям с небольшим доходом и вряд ли можно будет распространять выводы, полученные для полных наблюдений, на тех респондентов, которые не дали ответа на вопрос о возрасте. Как бы мы ни старались, из данных по бедным никак не извлечем данные по богатым. Подчеркнем, что мы говорим о ситуации, когда речь идет о точном восстановлении пропущенного значения возраста, а лишь такого восстановления этого значения, которое обеспечило бы нам сохранение соответствующих распределений. В данном случае сохранение распределения невозможно.

7. Влияние внешних факторов на возникновение и интерпретацию пропусков в анкетных данных

Практика показывает, что исследователи нередко неверно трактуют понятие случайности-неслучайности пропусков, и происходит это из-за неправильного понимания роли внешних факторов в формировании случайности. Рассмотрим несколько ситуаций (не буденм разделять «полную» и «простую» случайность.

Иногда социолог, натолкнувшись на необходимость определения неслучайности пропусков, утверждает, что пропуски неслучайны, если вероятность пропуска зависит от ненаблюдаемых факторов (т.е. от факторов, не охватываемых вопросами анкеты, от внешних факторов). Такое определение не выводит нас на выбор одного их тех общепринятых трех способов «борьбы» с пропусками, о которых шла речь во втором абзаце п.2. Когда мы говорим об этих способах, важна не зависимость от внешних факторов, а независимость от внутренних. Любой пропуск (как и любое другое явление) от чего-то всегда зависит. И, наверное, всегда можно говорить о зависимости от какого-либо внешнего фактора. Так что же, случайных пропусков у нас не бывает? Предположение, что пропуск от чего-то зависит, не делает его неслучайным. Как мы отмечали, случайность – понятие относительное. И нас интересует только та случайность, которая связывается с

возможностью рассмотрения случайных пропусков как значений случайных величин второго типа, с возможностью самого построения таких случайных величин (т.е. построения соответствующих распределений). Именно наличие или отсутствие такой возможности в сочетании с независимостью вероятности возникновения пропуска от внутренних факторов позволяет или не позволяет нам «ликвидировать» пропуски путем взвешивания выборки, либо искусственного заполнения пропусков. Мы определяем полную случайность возникновения пропуска при ответе респондента на какой-то вопрос в анкете как независимость соответствующей случайной величины второго типа от внутренних факторов. Внешние факторы здесь не при чем, хотя, конечно, пропуски вполне могут ими объясняться. А неслучайность возникает, когда мы не можем определить отвечающую рассматриваемому признаку случайную величину второго типа из-за того, что она зависит от случайной величины первого типа для того же признака.

Следующее критикуемое нами соображение сводится к тому, что при определенных условиях внешние факторы превращают случайные факторы в неслучайные. Имеется в виду некорректное использование часто упоминающегося в методической социологической литературе рассуждения о том, что ошибка при измерении мнения респондента о чем-либо может быть случайной, если используется адекватный инструмент измерения, и неслучайной (систематической), если инструмент измерения приводит к систематической ошибке в ответах респондентов. Конечно, изучение (и исключение) систематических ошибок при измерении - очень важный вопрос для социолога. Но это не имеет никакого отношения к интересующей нас случайности (неслучайности) неответа.

Возникновение методической ошибки может повысить или понизить вероятность неответа, но пропуск остается случайным, если он не зависит от внутренних факторов, и продолжает быть неслучайным, если вероятность неответа зависит от истинного значения признака для респондента.

Скажем, если интервьюер произведёт устрашающее впечатление на респондентов, и те станут реже отвечать на вопросы, это не повлечет перехода от случайности к неслучайности. Случайные пропуски для какого-либо вопроса вполне могут остаться случайными, просто изменится распределение соответствующей случайной величины второго типа: увеличится вероятность неответа. Но она вполне может остаться не зависимой от того, какая градация рассматриваемого признака в действительности респонденту соответствует.

Некоторыми из внешних факторов (например, методическими аспектами организации исследования) социолог может управлять. Естественно, если появление пропусков можно предотвратить, это надо делать. Но такое управление совсем не обязательно приведет к тому, что пропуски из случайных превратятся в неслучайные. Анализ влияния внешних факторов, обусловливающих возникновение неответов, может помочь исследователю разобраться со степенью случайности пропуска. Скажем, мы можем выявить, что вероятность ответа на какой-либо вопрос анкеты зависит от ответов на другие вопросы (например, показать, что действительно женщины более склонны скрывать свой возраст, чем мужчины) и тем самым доказать, что мы имеем дело с просто случайными пропусками. Можем выявить побуждают ли респондента какие-либо социальные причины дифференцированно относиться к разным значениям рассматриваемого признака и тем самым доказать, что имеем дело с неслучайными признаками и т.д.

8. Связь проблемы случайности пропусков с другими социологическими проблемами

Сделаем несколько замечаний о связи всего сказанного выше насчет способов решения проблемы пропущенных данных с некоторыми другими, не часто упоминающимися в литературе, методическими социологическими положениями.

Понятие генеральной совокупности

Понятие генеральной совокупности – это очень «темное» понятие для социолога. Далеко не всегда исследователю бывает понятно, что она из себя представляет. Все, сказанное о генеральной совокупности выше, может использоваться социологом и в таких задачах, которые лежат вне проблемы пропущенных данных. Подведем итог сказанному выше о понимании генеральной совокупности

Генеральная совокупность при любом многомерном статистическом анализе отождествляется с осмысленным для неё набором случайных величин, отвечающих отдельным признакам и отражающих их распределения (это – известный подход к пониманию генеральной совокупности [14, с.195]). Мы назвали их случайными величинами первого типа. В случае рассмотрения пропусков мы в понятие генеральной совокупности включаем еще один набор случайных величин: каждому признаку ставим в соответствие дихотомическую случайную величину второго типа, с которой связываем понятие случайности пропуска при ответе на рассматриваемый вопрос. И при использовании всех методов устранения пропусков мы стремились к

главной цели – сохранить все определяющие генеральную совокупность распределения.

Однородность изучаемой совокупности объектов

С помощью использования описанного понятия генеральной совокупности решается проблема однородности изучаемой совокупности объектов – одна из самых главных проблем, решение которой требуется для того, чтобы измерение и анализ данных давали бы адекватные результаты. О сути проблемы см.. например, [15]. Совокупность считается однородной, если она отвечает генеральной совокупности, понимаемой описанным выше образом. Другими словами, понятие однородности совокупности отождествляется с осмысленностью на ней какого-то набора случайных величин (и, как следствие, тех статистических закономерностей, которые можно найти путем их анализа). Подобное определение однородности известно из литературы (ссылки можно найти в [15]). Таким образом, понятие генеральной совокупности и понятие однородной совокупности фактически отождествляются.

Проблема однородности, как мы говорили в п.4, связана с проблемой случайности и по другой причине: вероятность пропуска может быть разной для разных категорий респондентов. Стремление к однородности порождает «просто» случайные и неслучайные пропуски. В данном случае однородность совокупности определяется через осмысленность для неё случайной величины, а осмысленность случайной величины – через однородность совокупности.

Переход от признака к его отдельным альтернативам

Мы уже неоднократно отмечали [16], что в социологии понятие признака часто имеет чисто номиналистский смысл: его удобно использовать для сбора информации, но при изучении социальных закономерностей оно иногда перестает работать, поскольку интересующее исследователя явление может объясняться тем, что признак принимает какое-либо конкретное значение, да еще, может быть, не само по себе, а в сочетании со значениями других признаков. В таких случаях говорят о поиске взаимодействий, т.е. сочетаний значений рассматриваемых признаков, детерминирующих то или иное поведение респондента. О роли признака в социологии, об относительной автономии отдельных градаций рассматриваемых признаков и о часто возникающей в социологии необходимости перехода от рассмотрения признака целиком к его отдельным альтернативам и к сочетаниям значений разных признаков (взаимодействий) мы говорили в [16, 17]. При рассмотрении проблемы пропусков мы сталкиваемся с той же ситуацией. Имеется в виду

возможность перехода от неслучайности пропусков к случайности. Кратко поясним это.

При определении неслучайных пропусков как раз и «работают» отдельные значения рассматриваемого признака. Рассмотрим пример, отраженный в таблице 1 Для всего признака целиком функция «ответ-неответ» явно не может рассматриваться как случайная величина. Слишком разбросаны доли неответивших на вопрос о возрасте среди людей разного возраста[8]. Единое вероятностное распределение не существует. Но при этом вполне может быть осмыслена вероятность того, что, скажем, человек старше 60 лет не дает ответа о своем возрасте с вероятностью 30%. И если мы дихотомизируем признак «возраст» (т.е., в нашем случае, превратим его в 4 полученных известным образом дихотомических признака) , то возникнет возможность каждое из получившихся дихотомических частотных распределений рассматривать как выборочное представление случайной величины второго типа и соответствующие пропуски считать случайными. Случайность возникла, когда мы от признака как целого перешли к отдельным его градациям. И это еще раз подтверждает относительность понятия случайности.

Литература

1. *Толстова Ю.Н.* Вероятностные и невероятностные модели порождения данных в социологии // Математическое моделирование социальных процессов. Вып. 12-13. М.: Спутник+, 2012. С.139-153.

2. *Platek R.* Causes of Incomple Data, Adjustments and Effects//Survey Methodology, Statistics Canada 1980, No.6, pp.93-132.

3. *Sande I.* Imputation in Surveys: Coping with Reality// The American Statistican,1982,Vol.36, No.3,Part 1, pp.145-152.

4. *Rubin D.B.* Multiple Imputation for Nonresponse in Surveys. NeyYerk: Willey, 1987.

5. *Little R. J. A.* A test of missing completely at random for multivariate data with missing values // Journal of the American Statistical Association. 1988. No. 83. P. 1198–1202.

6. *Esser H.* Determinanten des Interviewer–und Befragtenverhaltens: Probleme der theoretischenErklarung und empirischenUndersuchung von

[8] В идеале требуется определять статистическую значимость отклонения распределения, отраженного в этих долях, от равномерного. Если статистическую гипотезу о совпадении распределений имеет смысл принять, то признак логично считать случайным, даже если упомянутые доли отличаются друг от друга.

Interviewreffekten // Allgemeine Bevolkerungsumfrage der Sozialwissenschaften. Frankfurt, 1984. pp.314-336.

7. *Daubler T.* Nonresponseanalysen der Stichprobe F des SOEP. Berlin: DIW, 2002.P. 7-25.

8. *Hill D., WillisR.J.* Reducing Panel Attrition: A Search for Effective Policy Instruments // Journal of Human Resources 2001. Vol. 36. No.3. P. 416-438.

9. *Schrapler J.P.* Respondent Behavoir in Pannel Studies. A Case Study of the German Socio-Economic Panel (GSOEP) // DIW Discusion Paper . 2001. No.244. P.257-269.

10. *Nicoletti C., Peracchi F.* The Effects of Income Imputation on Microanalyses: Evidence from the European Community Household Panel // Journal of the Royal Statistical Society.2006. Series A (Statistics in Society).Vol. 169. No. 3. P. 625-646.

11. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // Computer Modelling&New Technologies. 2002. №. 6. С.49- 61.

12. Titterington D. M., Sedransk J. Matching and Linear Regression Adjustment in Imputation and Observational Studies // Sankhya: The Indian Journal of Statistics. 1986. Series B. Vol. 48. No. 3. P.347-367.

13. *Толстова Ю.Н.* Измерение в социологии. М.: ИДУ, 2009. – 292 с.

14. *Айвазян С.*А., Мхитарян В.С. Теория вероятностей и прикладная статистика. М.: Юнити, 2001.-656 с.

15. *Толстова Ю.Н.* Обеспечение однородности исходных данных в процессе применения математических методов // Социс, 1986, №3. С. 149-154.

16. *Толстова Ю.Н.* Анализ социологических данных. М.: Научный мир, 2000.- 350 с.

17. *Толстова Ю.Н.* Роль понятия признака при сборе и анализе социологических данных // Математическое моделирование социальных процессов. Вып. 12-13. М.: Спутник+, 2012. С.154-175.

В.А.Шведовский

Социологический факультет МГУ им. М.В.Ломоносова

СОЦИАЛЬНЫЕПРОГНОЗЫНАТОПОЛОГИЧЕСКИХДВОЙНЫХ ЦЕПЯХМАРКОВА[1]

1. Введение

1.1. Предварительные замечания

В социальных науках давно утвердился социальный прогноз, опирающийся на аппарат марковских цепей: это и прогнозирование социальной мобильности, рождаемости и численности различных социальных групп и т.д. В данной работе (вместе с рядом методологических посылок) анонсируется методика регулярных построений социальных прогнозов для социальных процессов, детерминируемых устойчивым образом жизни больших социальныхобщностей, репрезентируемых основным рядом типичных личностей этого социума.

В работе [11] на основании данных социологического исследования РГСУ-2010 при использовании аппарата цепей Маркова был построен прогноз численностей основных групп иммигрантов в Москву до 2015 года. В процессе применения аппарата цепей Маркова было установлено, что прогнозный результат особенно эффективен, если весь социум, для которого строится прогноз, предварительно разбивается в общностно-групповом аспекте на узнаваемые социальные группы – по критическому для исследуемого социального процесса признаку, для которых также естественно возникают отвечающие им типичные личности. Так при исследовании рождаемости иммигрантов Москвы органично определились, по своему демографическому поведению три группы: 1) тех, кто приехал с целью ассимилироваться с московским социумом, 2) тех, кто намерен не ассимилироваться и 3) тех, кто не определился с этим. Каждой из этих групп сопоставлялся соответствующий этой цели образ жизни, а, т.е. и отвечающий ему образ мыслей, что позволяло репрезентировать эти характерные социальные группы соответствующим им спектром типичных личностей. Это восходит к известной методике построения «социальных портретов». Таким образом, выдвигается методический приём замены изучения поведения социальных групп или общностей, детерминированного их траекториями по сферам образа жизни, на изучение траекторий между образами сфер жизнедеятельности - ячейками

[1] Работа выполнена при поддержке РФФИ (проекты 10-01-00332-а, 12-06-00205-а) и РГНФ (проект 12-03-00431)