Зангиева И. К., Москва

К вопросу о заполнении пропусков в социологических данных

Аннотация Статья посвящена различным аспектам заполнения пропусков в данных. Описываются основные разновидности неполной социологической информации: недостижимые и неполные наблюдения, созданные пропуски. Анализируется связь между причинами неответов на вопросы и степенью случайности порождаемых ими пропусков, определяющей допустимость их заполнения.

Ключевые слова: неполная информация, неполные наблюдения, отдельные пропуски, заполнение пропусков, неответы на вопросы, степень случайности, заполнение пропусков

Виды неполной информации

Начать следует с фиксации объекта исследования-основного фокуса данной статьи. В ней речь идет, прежде всего, об отдельных пропусках в данных, соответствующих неполным наблюдениям. Отдельные пропуски в данных являются частным случаем неполной информации, наряду с недостижимыми наблюдениями и сознательно созданными пропусками. Кратко охарактеризуем каждый из них.

Цель полевого этапа любого эмпирического социологического исследования – собрать максимальное количество релевантных данных, то есть получить ответы на все вопросы (в ситуации опроса) от всех запланированных респондентов. Если респондент отвечает на все поставленные вопросы ему соответствует полное наблюдение (full response). Задача полевого этапа - максимизация количества полных наблюдений. В идеале, который практически недостижим, их должно быть 100%.

В реальности исследователь всегда имеет дело с неполной информацией. Всегда есть респонденты, которых не удалось опросить и респонденты, которые будучи опрошенными, не ответили на некоторые заданные им вопросы. В первом случае возникают недостижимые наблюдения (unit – nonresponse), во втором - неполные наблюдения (отдельные пропуски-item /partial nonresponse). Неполные наблюдения, для которых известна только

часть информации, являются промежуточным звеном между полными наблюдениями (известная вся информация) и недостижимыми наблюдениями (нет информации вообще).

Недостижимые наблюдения (недостижимость респондентов)

В отечественной литературе в качестве синонима недостижимости часто используется понятие труднодоступности. Труднодоступными считаются респонденты, которые не могут (длительные командировки, болезни и т. д.), не хотят принять участие в опросе, которых трудно застать дома или невозможно опросить по причине того, что они проживают на отдаленных территориях [1; 4].Таким образом, недостижимыми являются респонденты, которых не удалось опросить в принципе.

Недостижимость в массовых социологических опросах представляет собой серьезную проблем. Наличие труднодоступных единиц наблюдения является источником систематических ошибок, т. к. недостижимые респонденты могут существенно отличаться от тех, кто в итоге принял участие в исследовании и ответил на вопросы [3, с. 59-61].

Масштабы недостижимости часто используются в роли индикаторов качества проведенного исследования и, прежде всего, реализации полевого этапа: «Высокий процент откликов стал синонимом эффективной и высококачественной организации исследования» [8, с .25].

Неполные наблюдения (отдельные пропуски)

Далее понятия «неполные наблюдения» и «отдельные пропуски» будут использоваться как синонимы.

Отдельные пропуски в данных можно разделить на два вида: реальные пропуски и вынужденные пропуски.

Реальные отдельные пропуски возникают когда, несмотря на все усилия исследователя или интервьюера (анкетера), респондент не отвечает на некоторые вопросы.

Вынужденные отдельные пропуски возникают в результате чистки массива, осуществляемой по завершении сбора и ввода данных. При чистке массива удаляются нереалистичные, заведомо ложные, нарушающие логику варианты ответа. Последние имеют место, если на один из вопросов респондент дает ответ, противоположный другим ответам на взаимосвязанные вопросы, нарушая тем самым всю логику последовательности [10, с. 145]. В литературе вынужденные пропуски так же называют искусственными.

Можно привести следующие примеры заведомо ложных значений. Подросток в качестве уровня образования указывает «кандидат наук», человек без определенного места жительства указывает площадь квартиры,

в которой якобы проживает на данный момент. Чтобы заведомо не вносить ложную информацию в данные, эти ответы из массива будут удалены, в результате чего возникнут искусственные, «артефактные» пропуски, или просто артефакты. [2].

Следует отметить, что наличие пропусков в данных наносит существенный урон качеству исследовательских результатов из-за:

- искажения распределений признаков (в некоторых случаях - возникновению систематических смещений);
- снижения статистической мощности результатов анализа данных в силу сокращения объема выборки;
- перехода порядковых шкал в частично упорядоченные;
- перехода непрерывных шкал в дискретные.

Сознательно созданные пропуски

Ограниченность ресурсов накладывает ограничения на стоимость, время проведения проекта и на количество вопросов, и, следовательно, тем, которые включаются в инструментарий. В некоторой степени уменьшить влияние данных факторов удается за счёт разбиения анкеты на несколько частей, предназначенных для различных групп респондентов, в рамках одного панельного исследования с чередующимися темами. Опрос в рамках одной волны нескольких групп респондентов по разным анкетам позволяет сэкономить временные, финансовые ресурсы и при этом ослабить нагрузку на респондентов.

Такой способ раздельного сбора данных с последующим объединением (слиянием) данных получил название data fusion.

Пропуски по вопросам блока анкеты, отсутствующему в данной волне некоторая группа респондентов не опрашивалась, затем заполняются с помощью стандартных алгоритмов заполнения отдельных пропусков [5; 11; 12].

Следует разделять пропуски, сознательно созданные исследователем еще на этапе планирования исследования, и незапланированные вынужденные пропуски, о которых было сказано выше. Первые представляют собой пропуски, полученные в ходе заранее спланированного экспериментального дизайна исследования, вторые же имеют вынужденный характер и изначально запланированы не были.

Решение проблемы недостижимости определенных респондентов и отдельные аспекты data fusion представляют собой крайне перспективные и актуальные направления для самостоятельных исследований и разработок, поэтому в данной работе мы не будем их далее рассматривать, а сосредоточимся только на работе с отдельными пропусками в данных.

Далее перейдем к работе с отдельными пропусками в данных. Существует 3 основных подхода к работе с ними уже после сбора данных: удаление неполных наблюдений, взвешивание имеющихся наблюдений для

достижения запланированного объема выборки и искусственное заполнение пропусков. В данной статье мы будем говорить только о заполнении пропусков, так как этот подход наиболее распространён в современной исследовательской практике и методической литературе и представляется наиболее перспективным.

Заполнение пропусков как центральный подход к работе с пропусками

В пользу актуальности заполнения пропусков для современной исследовательской практики и методической литературы говорит следующее. Статьи, посвященные различным аспектам заполнения пропусков, появляются в таких журналах, как Sociological Methods and Research (издательство Sage), Sociological Methodology (издательство Wiley), International Journal of Social Research Methodology (издательство Taylor & Francis). Первый из этих журналов занимает 6 место в рейтинге влиятельности 132 социологических журналов (данные Thomson Reuters, 2011).

Следует говорить именно об искусственном заполнении пропусков, так оно происходит уже «постфактум», с помощью математических или, что встречается значительно реже, логических процедур. Искусственности заполнения, в упомянутом смысле, можно было бы избежать, повторно обращаясь к каждому не ответившему на определенный вопрос респонденту с просьбой все-таки дать ответ на вопрос.

Заполнение пропусков имеет четыре основных сравнительных преимущества относительно удаления неполных наблюдений или взвешивания полных.

Во-первых, в отличие от взвешивания полных наблюдений, заполнение пропусков позволяет реально сохранить объем выборки на запланированном уровне.

Во-вторых, при заполнении пропусков, наряду с приращением новой информации, сохраняется вся известная информация, которая могла быть утеряна при удалении наблюдений с пропусками или взвешивании имеющихся.

В-третьих, в отличие от взвешивания полных наблюдений, заполнение пропусков не вызывает смещений по другим переменным, значения которых известны или в данный момент не восстанавливаются.

В-четвертых, после заполнения пропусков запланированный анализ данных может осуществляться в обычном режиме. Не нужно вводить дополнительных поправок, как например при взвешивании. Массив данных воспринимается и анализируется, как будто изначально от всех респондентов были получены ответы на все вопросы, и пропусков в данных не было в принципе.

Наряду с названными преимуществами заполнение пропусков как способ решения проблемы недостающей информации имеет несколько недостатков, которые нельзя не учитывать:

1. Использование для предсказания пропусков имеющихся данных может искажать общую структуру данных, которая смещается в сторону структуры только полных наблюдений.

2. Искусственное заполнение вносит в массив определенную долю (равную доле пропусков, в том случае если заполнялись все пропуски) искусственных данных.

Можно встретить точку зрения о неэтичности математического заполнения пропусков. Критики заполнения пропусков говорят о его неэтичности, обусловленной «вменением» не ответившим на вопрос респондентам «искусственных», рассчитанных или подобранных математическими способами значений (ответов), которые затем выдаются за истинные.

Нам данное соображение кажется в корне ошибочным. При заполнении пропусков не стоит задачи точного «угадывания» сокрытого ответа каждого не ответившего респондента. Задача заключается в восстановлении общего распределения изучаемого признака, искаженного наличием пропущенных значений. Здесь важно понимать, что заполнение пропусков математическими методами применяется в первую очередь в массовых количественных исследованиях, основанных на опросе большого числа респондентов. В силу «количественности» при анализе данных важно получить выводы обо всей изучаемой совокупности, а не о каждом ее отдельном представителе. Поэтому ответ каждого отдельного респондента как таковой значения не имеет. Исходя из этого, при заполнении пропусков происходит восстановление максимально достоверной статистической картины всей совокупности, а не «угадывание» ответа каждого не ответившего респондента или вменение, приписывание ему искусственно определенных значений [9].

Точность «угадывания» пропущенных значений используется как показатель эффективности заполнения пропусков в специальных методических экспериментах со смоделированными пропусками. В реальных же исследованиях точность подстановки нельзя оценить, так как истинное значение неизвестно.

При заполнении пропусков «физическое» приписывание ответов (значений переменных) – подстановка некоторых чисел на место каждого пропуска в массиве данных с помощью статистического пакета происходит только для того чтобы сделать возможной обработку данных с помощью традиционных методов анализа данных, предполагающих работу только с полными наблюдениями. По итогам заполнения пропусков ни в коем случае не говорится, что конкретный не ответивший на вопрос респондент на самом деле ответил согласно значению, подставленному на место имеющегося у него пропуска. Данное высказывание, действительно было бы неэтичным.

При заполнении пропусков этика не нарушается. Необходимо соблюдать этику при презентации и публикации результатов исследования. Этические соображения здесь требуют от исследователя в отчете по

результатам исследования или в любой другой публикации результатов указания на то, что имело место заполнение указанного количества пропусков конкретным способом (алгоритмом).

Однако, даже с соблюдением всех этических норм и только для получения обобщенных результатов обо всей совокупности в целом, заполнение пропусков допустимо и правомочно далеко не всегда. Допустимость заполнения пропусков определяется их характером, а именно степенью случайности.

Степень случайности пропусков как условие допустимости их заполнения

По степени случайности в литературе выделяют полностью случайные пропуски (missing completely at random– MCAR), случайные пропуски (missing at random– MAR) и неслучайные пропуски (not missing at random – NMAR).

Смысл каждого вида случайности можно пояснить на примере опроса следующим образом. Каждому вопросу в соответствие можно поставить случайную величину «ответ-неответ». Тогда степень случайности пропусков в ответах на конкретный вопрос определяется теми факторами, от которых зависит вероятность неответа респондентов на соответствующий вопрос (т. е. вид распределения дихотомической случайной величины «ответ-неответ», «привязанной» к каждому вопросу):

- при полной случайности пропусков вероятность неответа на вопрос не зависит ни от возможного ответа на данный вопрос, ни от ответов на другие вопросы. Распределение дихотомической величины «ответ-неответ» в данном случае одинаково при всех значениях данной переменной и при всех значениях остальных переменных.

- при случайности пропусков вероятность неответа не зависит от ответа на данный вопрос, но зависит от ответов на другие вопросы. Когда пропуски случайны распределение случайной величины «ответ-неответ» одинаково при всех значений рассматриваемого признака, но разное в группах, выделенных по значениям других рассматриваемых признаков.

- при неслучайности пропусков вероятность неответа на вопрос зависит от того, какой вариант ответа имеется в виду. Когда пропуски неслучайны и имеют систематический характер, распределение случайной величины «ответ-неответ» определить невозможно, так как оно разное для каждого значения рассматриваемой переменной [6, с.154-155].

При этом учитываются только факторы, отраженные в имеющейся у социолога информации о респондентах, т. е. в ответах на другие вопросы анкеты.

Степень случайности является математическим конструктом, оторванным от ситуации реального социологического исследования. Было бы полезно, помимо математического, найти и содержательное обоснование

допустимости заполнения пропусков. В качестве такого содержательного обоснования можно рассмотреть причины возникновения пропусков. Предположение о том, что пропуски каждой степени случайности порождаются определенными причинами, требует для своей проверки установления связи между причинами возникновения пропусков и степенью случайности порождаемых этими причинами пропусков.

В литературе выделяются три группы причин, по которым респонденты не отвечают на вопросы: психологические (различные характеристики личности респондента), социальные (особенности социальной ситуации и социального окружения в которых разворачивается ситуация опроса) и методические (различного рода ошибки, допущенные исследователем на этапе планирования исследования или интервьюером на этапе сбора данных) [7, с. 403-410].

Авторы, изучавшие основные причины возникновения пропусков (неответов респондентов на отдельные вопросы) не связывали причины неответов на вопросы со степенью случайности порождаемых ими пропусков. Аналогичное утверждение справедливо и для работ, посвященных изучению пропусков разной степени случайности: в этих работах практически не уделяется внимания причинам их возникновения. Другими словами, причины пропусков и их рассмотрение с точки зрения случайности в литературе рассматриваются раздельно. Это в определенном смысле естественно: первым аспектом фактически занимаются люди, решающие содержательные задачи (в нашем случае – социологи), вторым – математики. Попытаемся ликвидировать этот недостаток.

Говоря о выделенных выше типах причин возникновения пропусков нельзя установить жесткое соответствие между каждым типом причин и каждым типом пропусков по степени случайности.

В рамках каждой группы одни причины могут вызывать полностью случайные или случайные пропуски, а другие-не случайные.

На самом деле не всегда можно вычленить единственную причину, по которой респондент не ответил на вопрос. Процесс вопросно-ответной коммуникации иногда может быть подвержен влиянию нескольких причин одновременно. И, определение причин возникновения пропусков должно быть основано не столько на строгих доказательствах, сколько на опыте исследователя и его знаниях об особенностях темы и объекта исследованиях.

Выводы о связи между причинами возникновения пропусков, типами порождаемых ими пропусков по степени случайности и допустимыми способами работы с последними резюмируются в следующей таблице (см. таблицу 1).

Таблица 1

Связь между причинами возникновения пропусков, их типами и допустимыми способами корректировки после сбора данных

<table>
<tr><th rowspan="3">Причины возникновения пропусков</th><th>Вероятность неответа на вопрос:</th><th>Степень случайности, пропусков</th><th>Допустимый способ корректировки после сбора данных</th></tr>
<tr><td rowspan="2">Не зависит от возможного ответа</td><td>Полностью случайные (MCAR)</td><td rowspan="2">Удаление
Взвешивание
Заполнение</td></tr>
<tr><td>Случайные (MAR)*</td></tr>
<tr><td>Зависит от возможного ответа</td><td>Неслучайные (NMAR)</td><td>Не поддаются</td></tr>
</table>

*Для случайных пропусков перед взвешиванием и заполнением необходимо разбиение выборки на части, внутри которых пропуски полностью случайны.

Выше было отмечено, что между общими типами причин и степенью случайности порождаемых ими пропусков установить однозначное соответствие нельзя. Но можно утверждать следующее. Для обоснования допустимости корректировки после сбора данных (удаления, взвешивания выборки или заполнения пропусков) пропусков в ответах на определенный вопрос необходимо определить возможные причины возникновения последних и проанализировать связь между этими причинами и вероятностью неответа. Неслучайные пропуски, исключающие возможность их ликвидации после сбора данных, возникают под влиянием социальных, психологических или методических причин, только если последние ставят вероятность неответа на вопрос в зависимость от самого возможного «истинного ответа» (значения характеристики, измеряемой данным вопросом, которое было бы получено в случае ответа).

Поэтому, если у исследователя действительно есть основания полагать, что может иметь место ситуация возникновения неслучайных пропусков, похожая на одну из описанных выше, корректировка пропусков должна заключаться в максимальном устранении причин, породивших эту неслучайность. Заполнять неслучайные пропуски даже с помощью самых сложных алгоритмов некорректно, так как при заполнении пропусков алгоритмы, так или иначе, используют имеющиеся данные. Но в случае неслучайных пропусков, респонденты, не ответившие на вопрос, отличаются от ответивших как по значениям рассматриваемого признака, так и по значениям других признаков. Поэтому, некорректно при заполнении неслучайных пропусков использовать имеющиеся данные с совершенно другим распределением.

Еще одна практическая рекомендация, которую можно сделать, основываясь на приведенных выше размышлениях и примерах, касается ситуации, когда исследователь имеет дело со случайными пропусками. При случайных пропусках, когда вероятность неответа зависит от значений другого признака, внутри групп выделенных по значениям этого «другого» признака присутствует свое распределение вероятности неответа на вопрос. Поэтому заполнением пропусков необходимо разбить совокупность на группы, внутри которых пропуски по данной переменной полностью случайны (случайная величина «ответ-неответ» имеет одинаковое распределение внутри группы), и заполнять пропуски внутри каждой группы в отдельности. При работе со случайными пропусками возникает проблема поиска признаков определяющих случайность пропусков. Определить эти признаки необходимо для того, чтобы разбив по их значениям выборку, добиться в каждой подвыборке полной случайности пропусков. В первую очередь, в качестве таковых признаков имеет смысл рассматривать объективные характеристики респондентов, смысл которых очевиден, понятен и слабо зависит от способа измерения. И, здесь важно понимать, что зафиксировать абсолютно все признаки, определяющие именно случайность пропусков по некоторой переменной нельзя, потому что круг имеющихся потенциальных признаков – факторов случайности ограничен только признаками, изучаемыми в данном исследовании. И, может сложиться ситуация, что относительно изучаемых в исследовании признаков пропуски могут быть полностью случайными, а относительно не рассматриваемых в нем признаков - случайными. То есть, выводы о полной случайности и случайности пропусков, и соответственно допустимости их заполнения, могут быть справедливы только с точностью до признаков, изучаемых в рамках данного конкретного исследования.

Когда пропуски полностью случайны и в совокупности имеет место одно распределение случайной величины «ответ-неответ» заполнение пропусков правомочно в полной мере, так как респонденты, не ответившие на вопрос, не отличаются от ответивших ни по значениям рассматриваемого признака ни по значениям других признаков, и использование при заполнении пропусков имеющихся данных при корректной реализации не внесет в структуру данных и результаты их анализа специфических смещений.

Библиографический список

1. Бутенко И.А. «Нет ответа». Анализ методической ситуации на страницах журнала «Public Opinion Quarterly» // Социологические исследования. 1986. № 4. С.118-122.
2. Зангиева И.К. Проблема пропусков в социологических данных: смысл и подходы к решению // Социология: 4М (методология, методы, математические модели). 2011. № 33. С.28-56.

3. Хайкин С.Р., Павлов Э.П. Как помочь интервьюеру (из опыта методических исследований) // Социологические исследования. 1992. №4. С.48-64.
4. Чурилов Н.Н. Труднодоступные единицы исследования – источник систематических ошибок // Социологические исследования. 1986. № 1. С.64-78.
5. Adamek J. Fusion: Combining data from separate sources // Marketing Research: A Magazine of Management and Applications. 1994. Vol.6. No. 3. P.48-56.
6. D.de Leeuw E., Hox J., Huisman M. Prevention and Treatment of Item Nonresponse // Journal of Official Statistics. 2003. Vol. 19. No.2. P.155-156.
7. Ferber R. Item Nonresponse in a Consumer Survey // Public Opinion Quarterly. 1966. Vol. 30. No. 3. P. 399-415.
8. Ineke A.L.S. The Hunt for the Last Respondent. Nonresponse in sample surveys. Hague: Social and Cultural Planning Office of the Netherlands, 2005. P.18-35.
9. Rubin D.B. Multiple Imputation for Nonresponse in Surveys. New York: Willey, 1987. P. 64-69.
10. Sande I. Imputation in Surveys: Coping with Reality // The American Statistican.1982. Vol.36. No.3. P.145-152.
11. Wagner K., Wedel M. Factor Analysis and Missing Data // Journal of Marketing Research. 2000. No.11. P. 490-498.
12. Wagner K., Wedel M. Statistical Data Fusion for Cross – Tabulation // Journal of Marketing Research. 1997. No.11. P. 485-497.