ПРАКТИКИ СБОРА И АНАЛИЗА ФОРМАЛИЗОВАННЫХ ДАННЫХ

И.К. Зангиева, А.Н. Сулейманова
(*Москва*)

ПОДХОДЫ К АГРЕГИРОВАНИЮ РЕЗУЛЬТАТОВ МНОЖЕСТВЕННОГО ЗАПОЛНЕНИЯ ПРОПУСКОВ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Проведен сравнительный анализ эффективности применения правила Рубина и усреднения подставленных значений как подходов к агрегированию результатов множественного заполнения частичных пропусков в зависимости от исследовательской ситуации. При помощи статистического эксперимента оценена эффективность указанных подходов в исследовательских ситуациях, описываемых долей пропусков в массиве, типом шкалы переменных и методом анализа данных, который предполагается использовать после заполнения пропусков: описательная статистика, поиск связи между двумя признаками и множественная линейная регрессия. Для каждой рассмотренной исследовательской ситуации сформулированы рекомендации по выбору подхода к агрегации результатов множественного заполнения пропусков.

Ключевые слова: пропуски в данных, частичные пропуски, множественное заполнение пропусков, правило Рубина, агрегирование подставленных значений, исследовательская ситуация

Ирина Казбековна Зангиева – кандидат социологических наук, старший преподаватель кафедры методов сбора и анализа социологической информации, департамент социологии, факультет социальных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: izangieva@hse.ru.
Анна Наильевна Сулейманова – студентка магистерской программы «Прикладные методы социального анализа рынков», Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: ansuleymanova@edu.hse.ru.

Постановка исследовательской задачи

Неполные наблюдения или частичные пропуски (*item nonresponse*) в опросных данных возникают, если в процессе заполнения анкеты респондент отказался или не смог ответить на некоторые вопросы. Фокус на неполных наблюдениях объясняется тем, что, в отличие от недостижимых наблюдений (*unit nonresponse*), их наличие возможно скорректировать уже после сбора данных [1, с. 29]. Возможность корректировки частичных пропусков на этапе анализа данных определяется степенью случайности пропуска. По этому критерию пропуски делятся на полностью случайные (MCAR), случайные (MAR) и неслучайные, систематические (MNAR) [2, p. 9]. К игнорируемым пропускам, которые можно скорректировать на этапе анализа данных без обязательного учета механизма порождения пропусков [3, p. 33], относятся MAR и MCAR, а MNAR – к неигнорируемым, устраняемым только на этапе сбора информации при помощи исправления анкеты, дополнительного инструктажа интервьюеров или повторного обращения к респондентам. Неигнорируемым пропуск можно назвать в том случае, если распределение признака с пропусками среди ответивших отличается от его распределения среди тех, кто по каким-то причинам ответа не дал [4, с. 159].

Существует три основных группы методов борьбы с игнорируемыми пропусками в данных: удаление неполных наблюдений, взвешивание полных наблюдений для искусственного достижения запланированного объема выборки и заполнение пропусков. Активно развивается метод множественного заполнения пропусков, разработанный Дональдом Рубином в 1987 г. [5]. Этот метод предполагает подстановку на место каждого пропуска не одного значения, как в случае более простых методов заполнения, а нескольких. В результате исследователь получает несколько полных массивов, затем анализирует каждый из них и агрегирует результаты с применением специфических формул, называемых правилом Ру-

бина. Многократная подстановка пропущенных значений позволяет ввести поправку на неопределенность ответа, который мог бы дать респондент.

Проводить один и тот же анализ несколько раз на каждом массиве, а затем объединять результаты – задача достаточно трудоемкая. Для ее упрощения сам автор методики предлагал включать в анализ все заполненные массивы одновременно [6, p. 299], однако такой метод никогда не тестировался эмпирически применительно ко множественному заполнению пропусков. Другие исследователи в своих попытках упростить работу с алгоритмом ограничивались каким-либо специфическим видом анализа (к примеру, методом «отбора подобного по вероятности» – *propensity score estimation* [7]) или не слишком распространенной исследовательской ситуацией (например, когда есть возможность опросить всю генеральную совокупность и отпадает необходимость учитывать выборочную дисперсию [8]). Таким образом, не существует теоретических или эмпирических доказательств существования эффективных альтернатив правилу Рубина для всех прочих исследовательских ситуаций. В качестве такой альтернативы мы предлагаем рассмотреть агрегирование не результатов анализа данных, полученных на отдельных массивах, а подставляемых на место пропусков значений переменной. Усреднение подставленных значений (УПЗ) ускоряет и облегчает работу с множественным заполнением пропусков и в определенных исследовательских ситуациях, возможно, может служить эффективной заменой правилу Рубина.

Еще один аргумент в пользу предлагаемого способа агрегирования состоит в том, что не во всех статистических пакетах одинаково успешно объединена функция множественного заполнения пропусков и другие виды анализа или верификации результатов. К примеру, если анализ данных производится при помощи пакета SPSS любой из существующих на сегодняшний день версий, а дизайн исследования предполагает использование процедуры бутстреп, то правило Рубина придется применять

вручную, поскольку SPSS не позволяет произвести бутстреппинг на заполненных массивах.

Попытаемся установить, существуют ли исследовательские ситуации, когда агрегирование результатов множественного заполнения пропусков при помощи усреднения подставленных значений будет более эффективно, чем агрегирование с применением правила Рубина. Под *исследовательской ситуацией* здесь подразумевается комбинация типа шкалы переменной с пропусками, доли пропусков в массиве и применяемого метода анализа данных. Результатом должен стать набор рекомендаций по выбору наиболее эффективного подхода – правила Рубина или усреднения подставленных значений – для охваченных исследовательских ситуаций.

Теоретические предпосылки

Принципиальное отличие множественного заполнения пропусков от прочих методов борьбы с пропусками состоит в том, что каждый пропуск заменяется рассчитанным значением несколько раз, в результате чего исследователь получает несколько полных массивов. При подстановке значений следует помнить, что результаты заполнения не являются реальными ответами респондентов, и при анализе необходимо учитывать неопределенность, порождаемую совместным распределением переменной с пропусками и соответствующего ей индикатора присутствия, а также самой моделью заполнения [9, p. 581]. Тот факт, что в каждом из полученных массивов подставленные значения существенно различаются, эмпирически доказывает существование этой неопределенности [1, с. 46]. В случае применения множественного заполнения пропусков, к выборочной («внутримассивной») дисперсии добавляется «межмассивная» дисперсия, которая и позволяет рассматривать набор подставленных вместо конкретного пропуска значений как выборку, позволяющую установить не

истинный ответ респондента, а интервал, в котором с некоторой вероятностью лежит этот ответ.

Сам алгоритм состоит из четырех последовательных шагов [10, с. 205].

1. Обследование пропусков. На этом шаге исследователь, во-первых, определяет, к какой шкале относится переменная, содержащая пропуски; во-вторых, проверяет пропуски на монотонность; в-третьих, устанавливает, присутствуют ли в массиве переменные, которые можно использовать для расчета значений, подставляемых на место пропуска.

2. Определение модели заполнения. Исследователь, исходя из полученной на первом шаге информации, выбирает подходящую модель и включает в нее отобранные переменные-предикторы. Модель заполнения нужна для того, чтобы создать вероятностное распределение на основании наблюдаемых значений переменной и ее связей с переменными-предикторами, из которого затем в случайном порядке будут извлекаться значения для подстановки.

3. Подстановка значений. Из распределения, построенного на предыдущем шаге, случайным образом извлекаются m наборов значений переменной, которые подставляются в неполный массив, в результате чего мы получаем m полных массивов.

4. Анализ данных и агрегирование результатов. Классический алгоритм предполагает проведение анализа данных на каждом из m массивов и вычисление агрегированных статистических параметров при помощи правила Рубина. Приведем формулы для расчетов.

1. Для расчёта показателя:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m} Q_j,$$

где $\overline{Q}$ – агрегированный показатель, m – количество массивов, а Q_j – значение показателя для массива j.

2. Для оценки агрегированной стандартной ошибки. Внутригрупповая дисперсия:

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U_j,$$

где $\overline{U}$ – агрегированная внутригрупповая дисперсия показателя, а U_j– дисперсия показателя для массива j, и межгрупповая дисперсия B:

$$B = \frac{1}{m-1}\sum_{j=1}^{m} (Q_j - \overline{Q})^2,$$

при помощи которых вычислялась общая дисперсия T:

$$T = \overline{U} + B\left(1 + \frac{1}{m}\right).$$

3. Для проверки гипотез – критическое значение t-статистики со степенями свободы, рассчитанными по формуле:

$$df = (m-1)\left(1 + \frac{m\overline{U}}{B(m+1)}\right)^2,$$

и t-эмпирическим, равным

$$t = \overline{Q}/\sqrt{T}$$ [5, p. 76].

Метод множественного заполнения пропусков достаточно комплексный и трудоемкий, поэтому неоднократно предпринимались попытки упростить работу с ним. К примеру, в статье «Pooling Multiple Imputations When the Sample Happens to Be the Population» [8] рассматривалась ситуация, когда выборка исследования представляет собой всю генеральную совокупность (например, для редких медицинских состояний). В этом случае классический алгоритм переоценивает дисперсию из-за предположения о бесконечном объеме генеральной совокупности, заложенного в правило Рубина. В результате доверительные интервалы оказываются шире, чем требуется, а оценка – менее точной. Авторы

предложили скорректированное правило Рубина, которое принимает во внимание только дисперсию, обусловленную механизмом возникновения пропуска, и игнорирует выборочную дисперсию. В исследовании Р. Митры и Дж. Рейтера [7] сравнивались два метода агрегирования применительно к задаче измерения эффектов обработки. В первом случае агрегирование производилось по правилу Рубина, а во втором – заполненные значения для всех *m* массивов усреднялись и анализ производился на одном полном массиве. Рассматриваемый Митрой и Рейтером метод анализа был достаточно специфическим, а работ, посвященных сравнению эффективности разных подходов к агрегированию результатов множественного заполнения пропусков применительно к более распространенным исследовательским ситуациям, насколько нам известно, не существует.

Дизайн исследования

Нами была предпринята попытка сравнения эффективности двух подходов к агрегированию результатов множественного заполнения пропусков. Первый – классический – агрегирование результатов анализа по правилу Рубина. Второй возможный подход – усреднение подставляемых значений. Иными словами, наша цель состояла в том, чтобы оценить эффективность этих двух подходов к агрегированию результатов множественного заполнения пропусков в зависимости от исследовательской ситуации.

Нами были рассмотрены *исследовательские ситуации*, где сочетаются три предпосылки: тип шкалы переменной, содержащей пропуски (номинальная, порядковая и интервальная), доля пропусков в массиве (рассмотрены случаи 10, 30 и 50% пропусков) и метод анализа данных (описательная статистика, поиск связи между переменными и линейная регрессия).

Сравнить эффективность того или иного подхода теоретически достаточно трудно, поэтому для первичного тестирования

предположений, на которое направлена данное исследование, мы будем использовать статистический эксперимент.

Основная гипотеза исследования заключалась в том, что существуют такие исследовательские ситуации, когда для агрегирования результатов множественного заполнения пропусков усреднение подставленных значений оказывается эффективнее, чем правило Рубина.

В качестве информационной базы исследования выступает подвыборка жителей России (2484 наблюдения) из массива данных шестой волны Европейского социального исследования, проведенного в 2012 г.

В рамках статистического эксперимента нами была использована процедура бутстреп для верификации выводов об эффективности каждого подхода. Эта процедура предполагает интервальное оценивание параметров при помощи извлечения большого количества псевдовыборок с возвращением из эталонного массива полных наблюдений. Используя распределение значений параметра, полученных на каждой из выборок, рассчитывается стандартная ошибка и строится доверительный интервал [11, p. 577]. По умолчанию в пакете SPSS извлекается 1000 выборок, однако в большинстве исследований их количество варьируется от 50 до 10 000: В. Шитиков и Г. Розенберг [12] приводят примеры извлечения 5000 выборок, а Б. Эфрон [13] рассматривает случаи извлечения 4000 и 10 000 выборок; У. Джейкоби и Д. Армстронг применяли бутстреп с извлечением всего 50 выборок, аргументируя выбор тем, что рассматривают только «удобные для анализа» распределения (*well-behaved distributions*) [14, p. 271]. Мы ограничились тремя наборами – 1000, 10 000 и 50 000 выборок, поскольку в качестве основного критерия для сравнения эффективности рассматриваемых подходов выступают отклонения доверительных интервалов, полученных на заполненных массивах. При этом важно заметить, что при применении процедуры бутстреп реальная доверительная вероятность, с которой строится интервал, зависит от количества

выборок. К примеру, при извлечении 1000 выборок истинная доверительная вероятность интервала окажется между 93,6 и 96,4% с вероятностью 95%, а при извлечении 10 000 он сузится до границ 94,6 и 95,4% на том же уровне вероятности [15, p. 50–51].

Анализ данных, заполнение пропусков и верификация результатов производились при помощи статистического пакета IBM SPSS Statistics версии 21, поскольку именно этот пакет в России наиболее широко используется для анализа количественной социологической информации [10, с. 206].

При проверке всех статистических гипотез и построении доверительных интервалов мы задавались 95-процентным уровнем доверительной вероятности. Более высокая точность в целом не оправдана, поскольку критериями сравнения выступают отклонения границ доверительных интервалов, а не сами доверительные интервалы.

Проведенный в рамках исследования эксперимент состоял из 8 последовательных этапов.

Этап 1: отбор переменных. Переменные, которые использовались нами для демонстрации рассматриваемых методов анализа данных.

1. Описательная статистика: по одной переменной, измеряемой в номинальной, порядковой и интервальной шкале.

2. Поиск связи между признаками (наличие связи подразумевает значимость коэффициента на 95-процентном уровне доверительной вероятности, а отсутствие, соответственно, его незначимость на том же уровне вероятности):

– две номинальные переменные, между которыми существует немонотонная статистическая связь, и две номинальные переменные, между которыми немонотонной связи нет;

– две порядковые переменные, между которыми существует монотонная статистическая связь, и две порядковые переменные, между которыми монотонная связь отсутствует;

– две интервальные переменные, между которыми существует линейная статистическая связь, и две интервальные переменные,

между которыми линейная связь отсутствует.

3. Для построения модели множественной линейной регрессии: одна интервальная переменная на роль зависимой, две интервальные переменные на роль предикторов с незначимыми на 95-процентном уровне доверительной вероятности регрессионными коэффициентами и две интервальные переменные на роль предикторов со значимыми на том же уровне вероятности регрессионными коэффициентами.

В итоге нами были отобраны 13 переменных: 3 номинальные, 4 порядковые и 6 интервальных.

Этап 2: формирование эталонного массива. Для получения эталонного массива из исходной базы Европейского социального исследования нами отобраны только полные наблюдения, т.е. не имеющие ни одного пропуска в отобранных нами на предыдущем шаге переменных. Процедура была выполнена при помощи команды:

Select if (not missing(Var_1) … and not missing(Var_{13})
Execute.,

где Var_1 … Var_{13} – оставленные в массиве переменные.

В результате были отобраны 613 полных наблюдений.

Этап 3: фиксация эталонных результатов анализа данных. На эталонном массиве реализуются все запланированные нами виды анализа и с использованием бутстрепа фиксируются следующие параметры:

– для описательной статистики: выборочные доли, стандартные ошибки и 95-процентные доверительные интервалы для долей (номинальная и порядковая переменная), среднее арифметическое и дисперсия (интервальная переменная);

– для поиска связи между двумя признаками: выборочное значение, значимость, стандартная ошибка и 95-процентный дове-

рительный интервал для коэффициента *V* Крамера[1], коэффициента ранговой корреляции Спирмена и коэффициента корреляции Пирсона в случаях наличия и отсутствия соответствующих связей[2];

– для множественной линейной регрессии: выборочные значения, значимость, стандартные ошибки и доверительные интервалы для константы, значимых и незначимых на 95-процентном уровне доверительной вероятности регрессионных коэффициентов.

Этап 4: внесение в массив искусственных пропусков. Для охвата большего количества возможных исследовательских ситуаций было принято решение создать три экспериментальных массива: с 10, 30 и 50% пропусков. Выбор таких долей был обусловлен тем, что 10% пропусков – это та доля, которой, с одной стороны, зачастую легче пренебречь, а с другой – достаточно существенна потеря информации; 50% пропусков – это тот максимум, после которого странным представляется восстановление большей части отсутствующей информации за счет меньшей; 30% пропусков представляют середину между условной минимальной (10%) и максимальной (50%) долей пропусков.

Полностью случайные пропуски вносились в эталонный массив следующим образом.

1. Эталонный массив был перенесен в приложение Excel и каждому наблюдению был присвоен идентификационный номер.

2. При помощи команды СЛЧИС к базе была добавлена новая переменная, присваивающая каждому наблюдению случайное число от 1 до 613. Наблюдения были отсортированы по новой пере-

[1] Здесь для поиска немонотонной связи мы используем коэффициент, основанный на хи-квадрате, а не сам хи-квадрат, поскольку расчет доверительного интервала для коэффициента хи-квадрат с помощью процедуры бутстреп в SPSS не производится. Коэффициент *V* Крамера был предпочтен коэффициенту Фишера, поскольку тестируемые таблицы сопряженности отличались от *табл*. 2 на 2 ячейки.

[2] Под «наличием связи» мы подразумеваем, что коэффициент значим на 95-процентном уровне доверительной вероятности.

менной, после чего из столбца, содержащего значения первой экспериментальной переменной, были удалены первые 10% значений (расчет количества наблюдений, которые необходимо удалить для создания необходимой доли пропусков, см. в *табл*. 2). Далее была создана еще одна переменная, случайным образом присваивающая наблюдениям числа от 1 до 613, наблюдения снова сортировались по этой переменной, и из второй экспериментальной переменной удалялись первые 10% значений. Этот шаг был повторен 13 раз (по числу экспериментальных переменных). Создание случайной нумерации для внесения пропусков в каждую экспериментальную переменную необходимо, чтобы пропуски содержались не в одних и тех же наблюдениях, поскольку в этом случае наблюдения, оказавшиеся в начале списка обратились бы в полные неответы, к которым неприменим метод множественного заполнения пропусков.

3. Полученный экспериментальный массив с 10% искусственных пропусков был отсортирован по идентификационному номеру респондента и перенесен обратно в SPSS. Вся процедура повторялась еще два раза для создания массивов с 30 и 50% искусственных пропусков.

По итогам данного этапа нами были получены три массива данных с 10, 30 и 50% полностью случайных пропусков.

Этап 5: заполнение пропусков в экспериментальных массивах. Искусственные пропуски заполняются с созданием пяти[1] массивов на каждый экспериментальный с пропусками.

В модель заполнения, автоматически выбираемую SPSS, мы ввели все имеющиеся в массиве переменные. Поскольку на каждую из этих переменных в массиве имеется как минимум одна, с которой наблюдается статистическая связь, введение дополнительных переменных не требуется.

[1] Количество массивов, создаваемых в процессе множественного заполнения пропусков, обычно составляет от 2 до 10, поскольку большое количество подстановок не дает существенного увеличения эффективности оценки [10, с. 206]

Поскольку модели заполнения для номинальных переменных включали слишком много параметров, мы внесли следующие корректировки.

1. Максимальное число разрешенных параметров для модели заполнения было увеличено со 100 до 500; данная коррекция не влияет на качество заполнения, а лишь увеличивает время выполнения команды.

2. Три из шести интервальных переменных были только заполнены, но не использовались в качестве предикторов для других моделей заполнения, поскольку создавали большое количество категорий для логистической регрессии, но сами по себе ввиду слабых или отсутствующих связей со многими переменными не несли пользы для расчетов значений для заполнения прочих переменных.

В качестве метода заполнения нами была выбрана полностью условная спецификация, поскольку структура пропущенных данных в экспериментальных массивах немонотонная, в качестве модели для заполнения количественных переменных – линейная регрессия, для дискретных переменных SPSS по умолчанию применяет логистическую регрессию. Результатом стали 15 массивов, состоящих из полных наблюдений с подставленными значениями.

Этап 6: анализ данных на отдельных массивах с заполненными пропусками и агрегирование с применением правила Рубина. На каждом из 15 массивов проводились те же операции, что и на эталонном массиве. Результаты анализа данных агрегировались вручную в приложении Excel с применением правила Рубина.

Этап 7: агрегирование при помощи усреднения пропущенных значений и анализ данных на усредненных массивах. На этом этапе используются 15 массивов, полученные на пятом этапе.

Для пяти массивов, содержащих по 10, 30 и 50% заполненных данных, мы производим усреднение по следующей схеме:

– подставленные значения в переменных, измеренных в номинальной шкале, «усредняются» при помощи моды;

– подставленные значения в переменных, измеренных в порядковой шкале, «усредняются» при помощи медианы;

– подставленные значения в переменных, измеренных в интервальной шкале, усредняются при помощи среднего арифметического.

Результатом этой процедуры стали три массива из полных наблюдений, на которых мы произвели анализ данных по той же схеме, что и на эталонном массиве.

Этап 8: сравнение результатов с эталонными. Для того чтобы сравнить эффективность подходов, мы используем два критерия, введенных в эксперименте схожей тематики [16, с. 45]: степень отклонения и устойчивость доверительных интервалов.

Первый критерий – это пересечение эталонного доверительного интервала для оценки параметра с доверительным интервалом оценки, полученной на каждом из заполненных массивов, образованных на этапах 7 и 8, в рамках конкретной исследовательской ситуации, которое мы будем оценивать при помощи степени отклонения доверительных интервалов (Δ):

$$\Delta = \frac{|x_e - x_n| + |y_e - y_n|}{y_e - x_e} \times 100\% ,$$

где x_e – нижняя граница эталонного доверительного интервала, x_n – нижняя граница доверительного интервала, полученная после заполнения пропусков, y_e – верхняя граница эталонного доверительного интервала, y_n – верхняя граница доверительного интервала, полученная после заполнения пропусков.

Коэффициент выражает отношение абсолютного отклонения доверительного интервала от эталонного к длине эталонного доверительного интервала. Критерий принимает значения от нуля (если доверительный интервал совпадает с эталонным) до бесконечности, следовательно, чем меньше значение показателя, тем эффективнее подход к агрегированию результатов множественного заполнения пропусков. Более эффективным мы будем считать

тот подход, для которого Δ окажется ниже не менее, чем на 5%, в обратном случае разницу можно списать на статистическую погрешность, в случае чего методы будут признаны одинаково эффективными.

Второй критерий – устойчивость доверительных интервалов для параметров при изменении количества выборок, создаваемых при помощи бутстрепа. Под устойчивостью подразумевается неизменность доверительных интервалов при разном количестве псевдовыборок, т. е. более устойчивым будет признан тот подход, с применением которого доверительные интервалы будут колебаться в меньших пределах.

Ограничения дизайна исследования

Использование реальных выборочных данных, имеющих неизвестную ошибку измерения (*estimation bias*), в качестве основы для статистического эксперимента может показаться спорным. Кроме того, спорным может показаться использование массива, полученного из исходной базы ESS путем удаления неполных наблюдений, в качестве эталонного. В результате такой «очистки» объем доступной выборки сократился в 4 раза, что, разумеется, дает читателю основания усомнится в надежности и релевантности результатов анализа данных, полученных на эталонном массиве.

Указанные ограничения экспериментальной базы затрагивают вопрос: на какую совокупность можно распространять полученные нами результаты анализа данных и сделанные на их основе выводы об эффективности подходов к агрегированию результатов множественного заполнения пропусков? В связи с этим необходимо отметить, что для нашего исследования «генеральной совокупностью» является не генеральная совокупность ESS, а эталонный массив, состоящий только из полных наблюдений. Мы стремились оценить, насколько точно после применения каждого из рассматриваемых подходов к агрегированию воспроизводятся

тенденции, существующие именно в эталонном массиве. Чтобы не получить сугубо артефактные результаты, дополнительно применялась процедура бутстрепа, представляющая собой частный случай метода генерации испытаний Монте-Карло и позволяющая генерализовать результаты эксперимента.

Указанных проблем можно было бы избежать, изначально использовав вместо реальной базы сгенерированную, а в качестве инструмента генерализации – генерирование многих подвыборок методом Монте-Карло. Проведение такого эксперимента на сгенерированных данных мы рассматриваем в качестве направления для дальнейшей работы над темой.

Результаты статистического эксперимента

Сравнение эффективности подходов к агрегированию результатов множественного заполнения пропусков на основании сравнения степеней отклонения доверительных интервалов и устойчивости доверительных интервалов осуществляется при помощи таблиц (пример см. в *табл*. 1). Для каждой рассчитанной доли приводится доверительный интервал, построенный с извлечением 1000, 10000 и 50000 псевдовыборок и применением правила Рубина или усреднения подставленных значений. Далее для наглядности мы будем пользоваться обобщенными схемами. Со значениями точечных оценок, доверительных интервалов и степеней отклонения для всех рассмотренных исследовательских ситуаций можно ознакомиться в *Приложении*.

Описательная статистика. На *рис. 1* представлены подходы к агрегации результатов множественного заполнения пропусков, признанные более эффективными в соответствующих исследовательских ситуациях, предполагающих описание одномерных распределений с 10, 30 и 50% подставленных значений.

Для номинальной переменной в случае 10% подставленных значений в массиве правило Рубина оказалось эффективнее или

Таблица 1

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ПРАВИЛА РУБИНА И УПЗ ДЛЯ ОЦЕНКИ ДОЛЕЙ ЗНАЧЕНИЙ НОМИНАЛЬНОЙ ПЕРЕМЕННОЙ В МАССИВЕ С 10% ЗАПОЛНЕННЫХ ЗНАЧЕНИЙ

Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка (%)	Граница ДИ		Δ, %
		Граница ДИ				нижняя	верхняя	
		верхняя	нижняя					
1000	Крупный город	38,3	46,5	Правило Рубина	40,8	36,9	44,7	39
				Усреднение	*41,6*	*37,8*	*45,7*	*16*
	Пригород крупного города	2,0	4,9	*Правило Рубина*	*6,4*	*4,5*	*8,3*	*203*
				Усреднение	6,7	4,7	8,6	221
	Небольшой город	30,3	37,7	*Правило Рубина*	*33*	*29,3*	*36,7*	*27*
				Усреднение	29,2	25,3	33,1	130
	Деревня	17,1	23,5	Правило Рубина	16,3	13,2	19,3	127
				Усреднение	*19,6*	*16,6*	*23*	*16*
10 000	Крупный город	38,3	46,2	Правило Рубина	40,8	36,9	44,7	37
				Усреднение	*41,6*	*37,7*	*45,5*	*16*
	Пригород крупного города	2,1	4,9	*Правило Рубина*	*6,4*	*4,5*	*8,3*	*207*
				Усреднение	6,7	4,7	8,6	225
	Небольшой город	30,3	37,8	*Правило Рубина*	*33*	*29,3*	*36,7*	*28*
				Усреднение	32,1	28,5	35,9	49
	Деревня	17,1	23,5	*Правило Рубина*	*19,8*	*16,6*	*23*	*16*
				Усреднение	19,6	16,5	22,8	20

Окончание табл. 1

Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка (%)	Граница ДИ		Δ, %
		Граница ДИ				нижняя	верхняя	
		верхняя	нижняя					
50 000	Крупный город	38,3	46,2	Правило Рубина	40,8	36,9	44,7	37
				Усреднение	*41,6*	*37,7*	*45,5*	*16*
	Пригород крупного города	2,1	4,9	*Правило Рубина*	*6,4*	*4,5*	*8,3*	*207*
				Усреднение	6,7	4,7	8,6	225
	Небольшой город	30,3	37,8	*Правило Рубина*	*33*	*29,3*	*36,7*	*28*
				Усреднение	32,1	28,5	35,9	49
	Деревня	17,1	23,5	*Правило Рубина*	*19,8*	*16,6*	*23*	*16*
				Усреднение	19,6	16,5	22,7	22

Примечание. Для моделирования исследовательской ситуации, подразумевающей описание номинальной переменной, была отобрана переменная «Тип населенного пункта, в котором проживает респондент». Курсивом в таблице выделен подход, признанный более эффективным для конкретной доли значения признака при конкретном количестве извлеченных псевдовыборок.

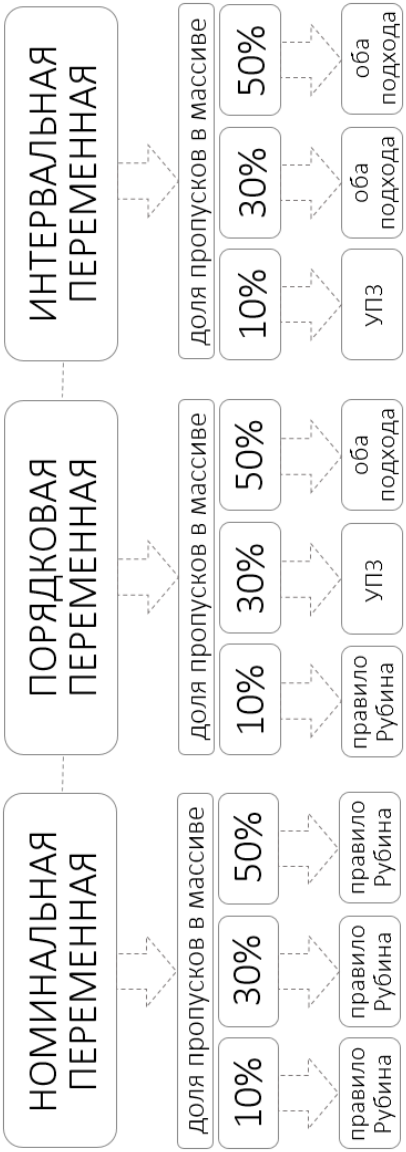

Рис. 1. **Результаты сравнения эффективности правила Рубина и усреднения подставленных значений для описательной статистики**

настолько же эффективно, как усреднение подставленных значений, и этот результат оставался неизменным при увеличении количества псевдовыборок. В массивах с 30% подставленных значений доверительный интервал, рассчитанный при помощи правила Рубина, оказывался ближе к эталонному, чем вычисленный на аналогичном усредненном массиве. Минимальная разница в степени отклонения составила 7, максимальная – 178%, причем результаты демонстрируют высокую устойчивость при увеличении количества извлекаемых выборок. Для массивов с 30% пропусков, таким образом, более эффективным оказалось правило Рубина. На массивах с 50% пропусков оба подхода показали достаточно низкую эффективность: для самой маленькой выборочной доли из четырех степень отклонения доверительного интервала достигает 1164%, в остальном же результаты повторяют полученные на массиве с 10% подставленных значений. Таким образом, при любом рассмотренном количестве пропусков в массиве для описания номинальной переменной более эффективным подходом оказывается применение правила Рубина.

В ситуации с описанием порядковой переменной для массива с 10% подставленных значений при любом количестве извлеченных выборок более эффективным оказалось правило Рубина. Отклонение в пользу усреднения для этого значения признака составило 11–12%, а в пользу правила Рубина для всех остальных долей – от 10 до 47%. Для массива с 30% пропусков более эффективным было признано усреднение подставленных значений. Применительно к массиву с 50% пропусков оба подхода имеют одинаковую эффективность при любом количестве извлеченных выборок. Мы можем сделать общий для описания порядковой переменной вывод, что в этой ситуации подход к агрегированию следует выбирать в зависимости от доли пропусков в массиве: для 10% пропусков более эффективным оказывается применение правила Рубина, для 30 – усреднение подставленных значений, а для 50% пропусков оба подхода одинаково эффективны.

Для вычисления среднего и дисперсии интервальной переменной результаты моделирования демонстрируют следующие тенденции:

– для массивов с 10% заполненных значений как для среднего, так и для дисперсии более эффективным подходом к агрегированию оказалось усреднение пропущенных значений;

– для массивов с 30% пропущенных значений для вычисления среднего более эффективным подходом к агрегированию оказалось усреднение подставленных значений, а для дисперсии – применение правила Рубина;

– для массивов с 50% заполненных значений для вычисления среднего более эффективным оказалось также усреднение пропущенных значений, а для дисперсии – правило Рубина вне зависимости от количества извлекаемых бутстрепом выборок.

Поиск связи между двумя признаками. Перейдем к сравнению эффективности подходов к агрегированию применительно к методам поиска связи между признаками начиная с коэффициента V Крамера, предназначенного для поиска немонотонной связи между признаками.

В ситуации отсутствия немонотонной связи между признаками более эффективным подходом оказывается усреднение подставленных значений для небольшого количества пропусков (10–30%), но в случае, если доля пропусков высока (50%), то разница между степенью отклонения для двух подходов составила меньше 5%, что можно списать на статистическую погрешность, значит – в этой исследовательской ситуации можно говорить об одинаковой эффективности обоих подходов. При увеличении количества извлекаемых бутстрепом выборок данные результаты оказались устойчивыми.

В ситуации же наличия связи между признаками усреднение подставленных значений более эффективен подход к агрегированию результатов заполнения для самой большой и самой маленькой долей пропусков, разница степени отклонения в пользу

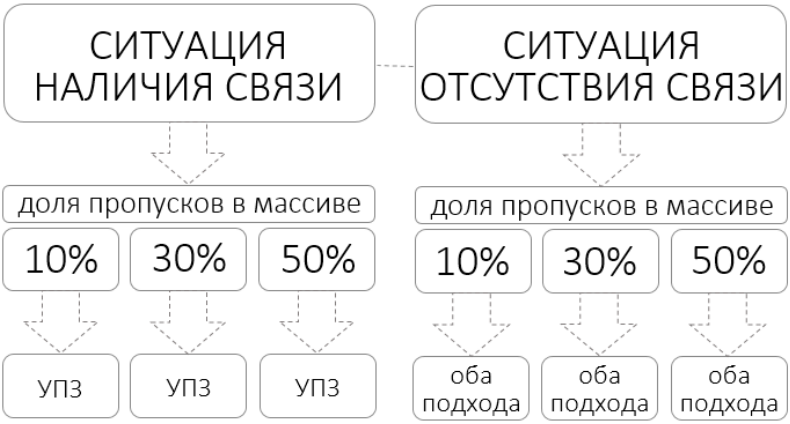

***Рис.* 2. Результаты сравнения эффективности правила Рубина и усреднения подставленных значений для вычисления коэффициента *V* Крамера**

усреднения составила от 7 до 53%. Для 30-процентной доли пропусков при любом количестве выборок разница между степенями отклонения для того или иного подхода составила от 0 до 4%, поэтому в данной исследовательской ситуации оба подхода демонстрируют одинаковую эффективность.

Таким образом, при вычислении коэффициента *V* Крамера усреднение подставленных значений предпочтительнее во всех исследовательских ситуациях, за исключением ситуации отсутствия немонотонной связи между признаками и большого количества пропусков в массиве, а также 30-процентных пропусков и наличия немонотонной связи между признаками: в этих случаях оба подхода одинаково эффективны.

Перейдем к сравнению результатов применительно к поиску монотонной связи между порядковыми признаками с использованием коэффициента ранговой корреляции Спирмена.

Применительно к ранговому коэффициенту Спирмена правило Рубина оказалось более эффективным подходом к агрегированию в

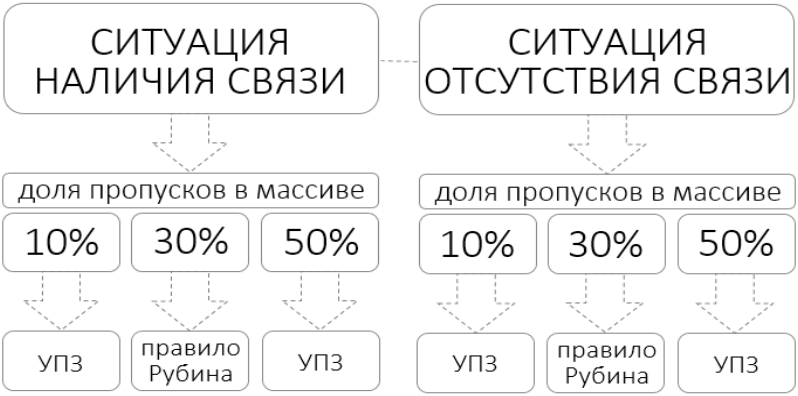

***Рис. 3.* Результаты сравнения эффективности правила Рубина и усреднения подставленных значений для коэффициента ранговой корреляции Спирмена**

том случае, если исследовательская ситуация характеризуется отсутствием монотонной связи между признаками и средним количеством пропусков в массиве (разница степеней отклонения в пользу правила Рубина составила от 64 до 72%). В случаях отсутствия связи и очень большого или очень маленького количества пропусков, усреднение подставленных значений оказывается более эффективным (разница составила от 11 до 44%). В данном случае также наблюдается устойчивость результатов вне зависимости от количества извлекаемых псевдовыборок.

В случае наличия монотонной связи между признаками ситуация идентична: более эффективным правило Рубина оказывается для среднего количества пропусков в массиве (разница степеней отклонения в пользу правила Рубина составляет 28–29%) и усреднение подставленных значений для очень большого и маленького количества пропусков (разница степеней отклонения составляет 20–167%).

Применительно к коэффициенту корреляции Пирсона более эффективным оказалось усреднение подставленных значений при всех рассмотренных долях пропусков в массиве в случае как

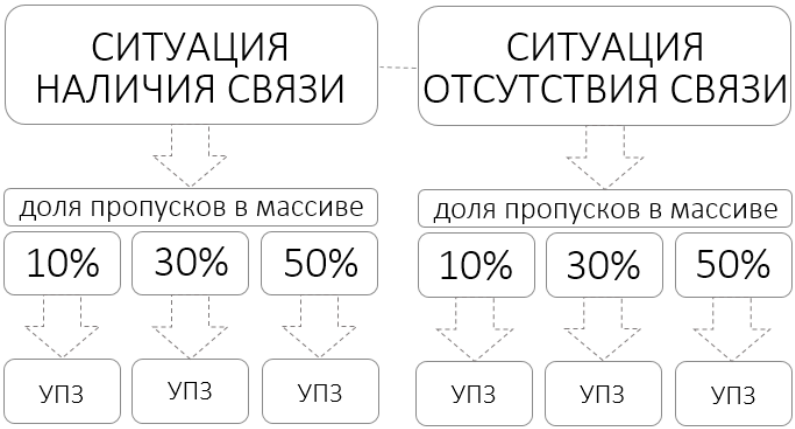

***Рис. 4*. Результаты сравнения эффективности правила Рубина и усреднения подставленных значений для вычисления коэффициента корреляции Пирсона**

наличия, так и отсутствия линейной связи между признаками. Разница степеней отклонения в пользу правила Рубина в случае отсутствия линейной связи составила от 38 до 72%, а в ситуации ее наличия – от 53 до 191%.

Множественная линейная регрессия. Перейдем к результатам сравнения эффективности подходов.

В случае, если в массиве присутствует небольшое число пропусков (10%), для константы границы доверительного интервала сильно колебались при изменении количества извлекаемых бутстрепом выборок для обоих подходов, однако в случаях 1 000 и 50 000 выборок этот интервал оказывался ближе к эталонному в том случае, если вычислялся на массиве с усредненными подставленными значениями. Для всех четырех коэффициентов регрессии при значимых и незначимых предикторах более эффективным оказывалось применение правила Рубина.

Результаты анализа данных, проведенные на массивах с 30% подставленных значений, в целом демонстрируют те же тенденции

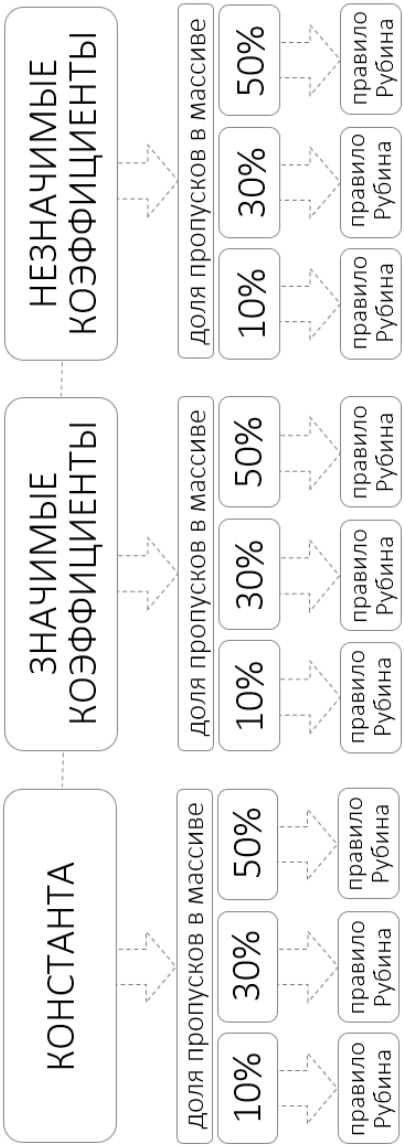

Рис. 5. **Результаты сравнения эффективности правила Рубина и усреднения подставленных значений для вычисления регрессионных коэффициентов и константы**

за одним исключением: для предиктора, измеренного в «истинной» интервальной шкале, со значимым регрессионным коэффициентом эффективнее оказалось усреднение подставленных значений. Учитывая первичный, пробный характер нашего исследования, мы не можем дать какой-либо теоретически подкрепленной интерпретации отличий именно для этого предиктора.

Результаты, полученные на массивах с 50% заполненных значений, аналогичны, однако для константы при любом количестве псевдовыборок более эффективным подходом оказалось правило Рубина. При этом для трех коэффициентов, доверительные интервалы для которых оказались ближе к эталону в случае применения усреднения подставленных значений, разница в степенях отклонения составила от 4 до 191%, а для коэффициента, который был оценен ближе к эталону при использовании правила Рубина – от 113 до 219%. Следовательно, во втором случае результаты оказались в среднем более чувствительны к изменению подхода к агрегированию. Поэтому с точки зрения частоты преобладания эффективности в данном случае более эффективным можно назвать усреднение пропущенных значений, а с точки зрения чувствительности результатов к изменению подхода – правило Рубина. Однако верхний предел разницы сопоставим в обоих случаях, поэтому ориентироваться стоит все же на частоту.

Подводя итог описанию результатов эксперимента, необходимо отметить, что наше основное предположение подтвердилось: среди рассмотренных исследовательских ситуаций нам действительно удалось выявить такие, для которых усреднение подставленных значений оказалось эффективнее правила Рубина. К таким ситуациям мы можем отнести описание порядковой переменной при умеренном количестве пропусков в массиве и описание интервальной переменной при небольшом количества пропусков; вычисление коэффициента *V* Крамера в случае, если исследователи предполагают наличие немонотонной связи между признаками и коэффициента Спирмена в случае небольшой или большой доли пропусков в массиве, а также коэффициента Пирсона в любой из рассмотренных ситуаций.

Заключение

При помощи статистического эксперимента мы оценили эффективность правила Рубина и усреднения пропущенных значений как подходов к агрегированию результатов множественного заполнения пропусков применительно к исследовательским ситуациям, характеризующимся различными долями пропусков в массиве, разными шкалами переменных и тремя распространенными в социологических исследованиях методами анализа данных (описательная статистика, поиск связи между двумя признаками и множественная линейная регрессия). На основании сравнения оценок эффективности подходов мы составили набор рекомендаций по выбору подхода к агрегированию результатов множественного заполнения пропусков для каждой из рассмотренных исследовательских ситуаций.

Описательная статистика

Для **описания номинальной переменной** при помощи долей значений признака в ситуации большого (до 50%), маленького (до 10%) и умеренного (до 30%) количества пропусков в массиве для агрегирования результатов множественного заполнения пропусков предпочтительно выбирать правило Рубина.

Для **описания порядковой переменной** при помощи долей значений признака для агрегирования результатов множественного заполнения пропусков предпочтительно выбирать:

а) в ситуации маленького (10%) количества пропусков в массиве – правило Рубина;

б) в ситуации умеренного (30%) количества пропусков в массиве – усреднение подставленных значений;

в) в ситуации большого (50%) количества пропусков в массиве оба подхода одинаково эффективны.

Для **описания интервальной переменной** при помощи среднего арифметического и дисперсии для агрегирования результатов множественного заполнения пропусков предпочтительно выбирать:

а) в ситуации небольшого (10%) количества пропусков в массиве – усреднение подставленных значений;

б) в ситуации умеренного (30%) количества пропусков в массиве – правило Рубина для среднего арифметического и усреднение подставленных значений для дисперсии;

в) в ситуации большого (50%) количества пропусков в массиве – усреднение подставленных значений для среднего арифметического и правило Рубина для дисперсии.

Поиск связи между двумя признаками

Для поиска немонотонной связи между двумя номинальными переменными с использованием **коэффициента V Крамера** для агрегирования результатов множественного заполнения пропусков предпочтительно выбирать:

а) в случае предположения о наличии немонотонной связи между признаками и любого (10–50%) количества пропусков в массиве – усреднение подставленных значений;

б) в случае предположения об отсутствии немонотонной связи между признаками и большого (50%), и небольшого (10%) или умеренного (30%) количества пропусков в массиве оба подхода одинаково эффективны.

Для поиска монотонной связи между двумя порядковыми переменными с использованием **коэффициента Спирмена** для агрегирования результатов множественного заполнения пропусков предпочтительно выбирать:

а) в случае предположения о наличии или отсутствии монотонной связи между двумя признаками и небольшого (10%) или большого (50%) количества пропусков в массиве – усреднение подставленных значений;

б) в случае предположения о наличии или отсутствии монотонной связи между двумя признаками и умеренного (30%) количества пропусков в массиве – правило Рубина.

Для поиска линейной связи между двумя интервальными переменными с использованием **коэффициента Пирсона** для агрегирования результатов множественного заполнения пропусков в случае предположения о наличии или отсутствии линейной связи и любого (10–50%) количества пропусков в массиве предпочтительно выбирать усреднение подставленных значений.

Множественная линейная регрессия

Для оценки константы и предположений о значимости или незначимости регрессионных коэффициентов во множественной линейной регрессии в ситуации любого (10–50%) количества пропусков в массиве для агрегирования результатов множественного заполнения пропусков предпочтительно выбирать правило Рубина.

Сфера применения результатов исследования ограничивается только при наличии полностью случайных пропусков. Кроме того, нами была рассмотрена лишь небольшая часть возможных исследовательских ситуаций: всего три метода анализа данных из очень широкого круга статистических методов, применяемых в социологии. В связи с этим делать теоретические или методические обобщения на основании данного исследования можно только применительно к рассмотренным исследовательским ситуациям или ситуациям, близким к ним. Основным результатом исследования стало экспериментальное доказательство того, что в ряде случаев более простой в осуществлении метод – усреднение подставленных значений при помощи соответствующей меры центральной тенденции – оказывается эффективнее правила Рубина.

Таким образом, нам удалось проложить новое направление для оптимизации применения множественного заполнения пропусков в зависимости от исследовательской ситуации. Дальнейшие исследования в данной области могут касаться следующих проблем:

– теоретическое обоснование адекватности применения усреднения подставленных значений для агрегирования результатов множественного заполнения пропусков;

– расширение круга экспериментально обоснованных рекомендаций по выбору подхода к агрегированию результатов множественного заполнения пропусков в различных исследовательских ситуациях;

– теоретическое обоснование эффективности усреднения пропущенных значений или применения правила Рубина в конкретных исследовательских ситуациях.

ЛИТЕРАТУРА

1. *Зангиева И.К.* Проблема пропусков в социологических данных: смысл и подходы к решению // Социология: методология, методы, математическое моделирование. 2011. № 33. С. 28–56.

2. *Brand J.P.L.* Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. Thesis Erasmus University Rotterdam, 1999.

3. *Jin H., Rubin D.* Public Schools versus Private Schools: Causal Inference with Partial Compliance // Journal of Educational and Behavioral Statistics. 2009. Vol. 34. No. 1. P. 24– 45.

4. *Зангиева И.К., Толстова Ю.Н.* Понятие случайности и проблема пропусков данных в социологии // Математическое моделирование социальных процессов / Под ред. А. Михайлова. М.: Социологический факультет МГУ, 2012. Вып. 14. Гл. 14. С. 146–165.

5. *Rubin D.* Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, 1987.

6. *Rubin D.* The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys // The American Statistician. 2004. Vol. 58. No. 4. P. 298–302.

7. *Mitra R., Reiter J.P.* A Comparison of Two Methods of Estimating Propensity Scores After Multiple Imputation // Statistical Methods in Medical Research. 2016. Vol. 25. Iss. 1. P. 188–204.

8. *Vink G., van Buuren S.* Pooling Multiple Imputations When the Sample Happens to Be the Population [online source] // Cornell University Library. 2014. URL: http://arxiv.org/abs/1409.8542 (date of access: May 3, 2016).

9. *Zhang P.* Multiple Imputation: Theory and Method // International Statistical Review. 2003. Vol. 71. No. 3. P. 581-592.

10. *Кутлалиев А.Х.* Метод множественного восстановления данных // Социологические методы в современной исследовательской практике: сборник статей, посвященный памяти первого декана факультета социологии НИУ ВШЭ А.О. Крыштановского [Электронный ресурс] / Под ред. О.А. Оберемко. М.: Издательский дом НИУ ВШЭ, 2011. С. 201–208.

11. *Kromrey J.D., Hines C.V.* Nonrandomly Missing Data in Multiple Regression: An Empirical Comparison of Common Missing-Data Treatments // Educational and Psychological Measurement. 2003. Vol. 54. P. 573–593.

12. *Шитиков В.К., Розенберг Г.С.* Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. Тольятти: Кассандра, 2013.

13. *Efron B.* Bayesian Inference and the Parametric Bootstrap // The Annals of Applied Statistics. 2012. Vol. 6. No. 4. P. 1971–1997.

14. *Jacoby W., Armstrong D.II* Bootstrap Confidence Regions for Multidimensional Scaling Solutions // American Journal of Political Science. 2014. Vol. 58. No. 1. P. 264–278.

15. *Manly B.* Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition. Chapman and Hall / CRC, 2006.

16. *Зангиева И.К., Тимонина Е.С.* Сравнение эффективности алгоритмов заполнения пропусков в данных в зависимости от используемого метода анализа // Мониторинг общественного мнения. 2014. № 1(119). С. 41–55.

Приложение

Таблица 1

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ПРАВИЛА РУБИНА И УСРЕДНЕНИЯ ПОДСТАВЛЕННЫХ ЗНАЧЕНИЙ ДЛЯ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Номинальная («тип населенного пункта, в котором проживает респондент»)	10	1000	Крупный город	38,3	46,5	Правило Рубина	40,8	36,9	44,7	39
						Усреднение	*41,6*	*37,8*	*45,7*	*16*
			Пригород крупного города	2,0	4,9	Правило Рубина	6,4	4,5	8,3	203
						Усреднение	*6,7*	*4,7*	*8,6*	*221*
			Небольшой город	30,3	37,7	*Правило Рубина*	*33*	*29,3*	*36,7*	*27*
						Усреднение	29,2	25,3	33,1	130
			Деревня	17,1	23,5	Правило Рубина	16,3	13,2	19,3	127
						Усреднение	*19,6*	*16,6*	*23*	*16*
		10 000	Крупный город	38,3	46,2	Правило Рубина	40,8	36,9	44,7	37
						Усреднение	*41,6*	*37,7*	*45,5*	*16*
			Пригород крупного города	2,1	4,9	*Правило Рубина*	*6,4*	*4,5*	*8,3*	*207*
						Усреднение	6,7	4,7	8,6	225
			Небольшой город	30,3	37,8	*Правило Рубина*	*33*	*29,3*	*36,7*	*28*
						Усреднение	32,1	28,5	35,9	49
			Деревня	17,1	23,5	*Правило Рубина*	*19,8*	*16,6*	*23*	*16*
						Усреднение	*19,6*	*16,5*	*22,8*	*20*

Продолжение табл. 1

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон			Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ					нижняя	верхняя	
				нижняя	верхняя						
Номинальная («тип населенного пункта, в котором проживает респондент»)	10	50 000	Крупный город	38,3	46,2		Правило Рубина	40,8	36,9	44,7	37
							Усреднение	*41,6*	*37,7*	*45,5*	*16*
			Пригород крупного города	2,1	4,9		*Правило Рубина*	*6,4*	*4,5*	*8,3*	*207*
							Усреднение	6,7	4,7	8,6	225
			Небольшой город	30,3	37,8		*Правило Рубина*	*33*	*29,3*	*36,7*	*28*
							Усреднение	32,1	28,5	35,9	49
			Деревня	17,1	23,5		*Правило Рубина*	*19,8*	*16,6*	*23*	*16*
							Усреднение	19,6	16,5	22,7	22
	30	1000	Крупный город	38,3	46,5		*Правило Рубина*	*39,9*	*36*	*43,7*	*62*
							Усреднение	39,5	35,6	43,2	73
			Пригород крупного города	2,0	4,9		*Правило Рубина*	*11,7*	*9,1*	*14,3*	*569*
							Усреднение	13,9	11,3	16,6	724
			Небольшой город	30,3	37,7		*Правило Рубина*	*30,3*	*26,7*	*34*	*99*
							Усреднение	29,2	25,3	33,1	130
			Деревня	17,1	23,5		*Правило Рубина*	*18,1*	*15,1*	*21,2*	*67*
							Усреднение	17,5	14,5	20,6	86
		10 000	Крупный город	38,3	46,2		*Правило Рубина*	*39,9*	*36*	*43,7*	*61*
							Усреднение	39,5	35,7	43,4	68
			Пригород крупного города	2,1	4,9		*Правило Рубина*	*11,7*	*9,1*	*14,3*	*586*
							Усреднение	13,9	11,1	16,6	739

Продолжение табл. 1

Тип перемешивания	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Номинальная («тип населенного пункта, в котором проживает респондент»)	30	10 000	Небольшой город	30,3	37,8	*Правило Рубина*	*30,3*	*26,7*	*34*	*99*
						Усреднение	29,2	25,6	32,8	129
			Деревня	17,1	23,5	*Правило Рубина*	*18,1*	*15,1*	*21,2*	*67*
						Усреднение	17,5	14,5	20,6	86
		50 000	Крупный город	38,3	46,2	*Правило Рубина*	*39,9*	*36*	*43,7*	*61*
						Усреднение	39,5	35,7	43,4	68
			Пригород крупного города	2,1	4,9	*Правило Рубина*	*11,7*	*9,1*	*14,3*	*586*
						Усреднение	13,9	11,3	16,6	746
			Небольшой город	30,3	37,8	*Правило Рубина*	*30,3*	*26,7*	*34*	*99*
						Усреднение	29,2	25,6	32,8	129
			Деревня	17,1	23,5	*Правило Рубина*	*18,1*	*15,1*	*21,2*	*67*
						Усреднение	17,5	14,5	20,6	86
	50	1000	Крупный город	38,3	46,5	Правило Рубина	35,4	31,6	39,2	171
						Усреднение	*35,9*	*32,1*	*39,5*	*161*
			Пригород крупного города	2,0	4,9	*Правило Рубина*	*17,1*	*14,1*	*20,1*	*941*
						Усреднение	19,7	16,6	23	1128
			Небольшой город	30,3	37,7	*Правило Рубина*	*29,5*	*25,9*	*33,2*	*120*
						Усреднение	27,6	24	31	176
			Деревня	17,1	23,5	*Правило Рубина*	*17,9*	*14,8*	*21*	*75*
						Усреднение	16,8	13,9	19,7	109

Продолжение табл. 1

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Номинальная («тип населенного пункта, в котором проживает респондент»)	50	10 000	Крупный город	38,3	46,2	Правило Рубина	35,4	31,6	39,2	173
						Усреднение	*35,9*	*32,1*	*39,6*	*162*
			Пригород крупного города	2,1	4,9	Правило Рубина	17,1	14,1	20,1	971
						Усреднение	19,7	16,6	23	1164
			Небольшой город	30,3	37,8	*Правило Рубина*	*29,5*	*25,9*	*33,2*	*120*
						Усреднение	27,6	24	31,2	172
			Деревня	17,1	23,5	*Правило Рубина*	*17,9*	*14,8*	*21*	*75*
						Усреднение	*16,8*	*13,9*	*19,7*	*109*
		50 000	Крупный город	38,3	46,2	Правило Рубина	35,4	31,6	39,2	173
						Усреднение	*35,9*	*32,1*	*39,6*	*162*
			Пригород крупного города	2,1	4,9	*Правило Рубина*	*17,1*	*14,1*	*20,1*	*971*
						Усреднение	19,7	16,6	23	1164
			Небольшой город	30,3	37,8	*Правило Рубина*	*29,5*	*25,9*	*33,2*	*120*
						Усреднение	27,6	24	31,2	172
			Деревня	17,1	23,5	*Правило Рубина*	*17,9*	*14,8*	*21*	*75*
						Усреднение	16,8	13,9	19,7	109

Продолжение табл. 1

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Порядковая («Заинтересованность в политике»)	10	1000	Очень заинтересован	14,0	19,7	Правило Рубина	17,5	14,5	20,6	25
						Усреднение	*17,1*	*14,4*	*20,1*	*14*
			Довольно заинтересован	38,5	46,3	*Правило Рубина*	*42,4*	*38,5*	*46,3*	*0*
						Усреднение	44	40,5	48	47
			Едва ли заинтересован	27,1	34,3	*Правило Рубина*	*29,8*	*26,2*	*33,4*	*25*
						Усреднение	29,4	25,9	32,8	38
			Совсем не заинтересован	7,7	12,6	*Правило Рубина*	*10,3*	*7,9*	*12,7*	*6*
						Усреднение	9,5	7,2	11,7	29
		10 000	Очень заинтересован	13,9	19,7	Правило Рубина	17,5	14,5	20,6	26
						Усреднение	*17,1*	*14,2*	*20,2*	*14*
			Довольно заинтересован	38,7	46,5	*Правило Рубина*	*42,6*	*38,7*	*46,5*	*0*
						Усреднение	44	40,1	48,1	38
			Едва ли заинтересован	27,1	34,4	*Правило Рубина*	*29,8*	*26,2*	*33,4*	*26*
						Усреднение	29,4	25,8	33	37
			Совсем не заинтересован	7,7	12,4	*Правило Рубина*	*10,3*	*7,9*	*12,7*	*11*
						Усреднение	9,5	7,2	11,9	21
		50 000	Очень заинтересован	13,9	19,7	Правило Рубина	17,5	14,5	20,6	26
						Усреднение	*17,1*	*14,2*	*20,2*	*14*
			Довольно заинтересован	38,7	46,5	*Правило Рубина*	*42,6*	*38,7*	*46,5*	*0*
						Усреднение	44	40,1	48,1	38

Продолжение табл. 1

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Порядковая («Заинтересованность в политике»)	10	50 000	Едва ли заинтересован	27,1	34,4	*Правило Рубина*	*29,8*	*26,2*	*33,4*	*26*
						Усреднение	29,4	25,8	33	37
			Совсем не заинтересован	7,7	12,4	*Правило Рубина*	*10,3*	*7,9*	*12,7*	*11*
						Усреднение	9,5	7,2	11,9	21
	30	1000	Очень заинтересован	14,0	19,7	Правило Рубина	39,5	35,7	43,4	796
						Усреднение	*15,7*	*12,7*	*18,8*	*39*
			Довольно заинтересован	38,5	46,3	Правило Рубина	39,5	35,7	43,4	73
						Усреднение	*44,9*	*40,8*	*48,6*	*59*
			Едва ли заинтересован	27,1	34,3	Правило Рубина	27,8	24,3	31,4	79
						Усреднение	*30,3*	*26,6*	*34,1*	*10*
			Совсем не заинтересован	7,7	12,6	Правило Рубина	15,7	12,8	18,6	227
						Усреднение	*9,1*	*7*	*11,3*	*41*
		10 000	Очень заинтересован	13,9	19,7	Правило Рубина	39,5	35,7	43,4	784
						Усреднение	*15,7*	*12,9*	*18,6*	*36*
			Довольно заинтересован	38,7	46,5	Правило Рубина	39,5	35,7	43,4	78
						Усреднение	*44,9*	*40,9*	*48,8*	*58*
			Едва ли заинтересован	27,1	34,4	Правило Рубина	27,8	24,3	31,4	79
						Усреднение	*30,3*	*26,8*	*34,1*	*8*
			Совсем не заинтересован	7,7	12,4	Правило Рубина	15,7	12,8	18,6	240
						Усреднение	*9,1*	*6,9*	*11,4*	*38*

Продолжение табл. 1

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Порядковая («Заинтересованность в политике»)	30	50 000	Очень заинтересован	13,9	19,7	Правило Рубина	39,5	35,7	43,4	784
						Усреднение	*15,7*	*12,9*	*18,6*	*36*
			Довольно заинтересован	38,7	46,5	Правило Рубина	39,5	35,7	43,4	78
						Усреднение	*44,9*	*40,9*	*48,8*	*58*
			Едва ли заинтересован	27,1	34,4	Правило Рубина	27,8	24,3	31,4	79
						Усреднение	*30,3*	*26,8*	*34*	*10*
			Совсем не заинтересован	7,7	12,4	Правило Рубина	15,7	12,8	18,6	240
						Усреднение	*9,1*	*6,9*	*11,4*	*38*
	50	1000	Очень заинтересован	14,0	19,7	Правило Рубина	19,1	16	22,2	79
						Усреднение	*17,1*	*14,2*	*20,1*	*11*
			Довольно заинтересован	38,5	46,3	*Правило Рубина*	*38,7*	*34,8*	*42,5*	*96*
						Усреднение	51,4	47	55,1	222
			Едва ли заинтересован	27,1	34,3	*Правило Рубина*	*25,9*	*22,4*	*29,3*	*135*
						Усреднение	25,4	22,2	29,2	139
			Совсем не заинтересован	7,7	12,6	Правило Рубина	16,4	13,5	19,3	255
						Усреднение	*6*	*4,2*	*8*	*165*
		10 000	Очень заинтересован	13,9	19,7	Правило Рубина	19,1	16	22,2	79
						Усреднение	*17,1*	*14,2*	*20,1*	*12*
			Довольно заинтересован	38,7	46,5	*Правило Рубина*	*38,7*	*34,8*	*42,5*	*101*
						Усреднение	*38,7*	*34,8*	*42,5*	*101*

Окончание табл. 1

Тип переменной	Доля пропусков в массиве, %	Извлеченные выборки	Значение признака	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Порядковая («Заинтересованность в политике»)	50	10 000	Едва ли заинтересован	27,1	34,4	*Правило Рубина*	*25,9*	*22,4*	*29,3*	*134*
						Усреднение	25,4	22	28,9	145
			Совсем не заинтересован	7,7	12,4	Правило Рубина	16,4	13,5	19,3	270
						Усреднение	*6*	*4,2*	*8*	*168*
		50 000	Очень заинтересован	13,9	19,7	Правило Рубина	19,1	16	22,2	79
						Усреднение	*17,1*	*14,2*	*20,1*	*12*
			Довольно заинтересован	38,7	46,5	*Правило Рубина*	*38,7*	*34,8*	*42,5*	*101*
						Усреднение	*38,7*	*34,8*	*42,5*	*101*
			Едва ли заинтересован	27,1	34,4	*Правило Рубина*	*25,9*	*22,4*	*29,3*	*134*
						Усреднение	25,4	22	28,9	145
			Совсем не заинтересован	7,7	12,4	Правило Рубина	16,4	13,5	19,3	270
						Усреднение	*6*	*4,2*	*8*	*168*

Таблица 2

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ПРАВИЛА РУБИНА И УСРЕДНЕНИЯ ПОДСТАВЛЕННЫХ ЗНАЧЕНИЙ ДЛЯ ОЦЕНКИ СРЕДНЕГО И ДИСПЕРСИИ ИНТЕРВАЛЬНОЙ ПЕРЕМЕННОЙ

Доля пропусков, %	Извлеченные выборки	Параметр	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			верхняя	нижняя					
10	1000	Среднее	12,89	13,36	Правило Рубина	13,07	13,04	13,10	87
					Усреднение	*13,07*	*12,85*	*13,30*	*21*
		Дисперсия	7,306	9,581	Правило Рубина	8,61	8,55	8,67	95
					Усреднение	*8,077*	*6,966*	*9,293*	*28*
	10 000	Среднее	12,90	13,36	Правило Рубина	13,07	13,04	13,10	87
					Усреднение	*13,07*	*12,85*	*13,29*	*26*
		Дисперсия	7,286	9,580	Правило Рубина	8,61	8,55	8,67	95
					Усреднение	*8,077*	*7,004*	*9,242*	*27*
	50 000	Среднее	12,90	13,36	Правило Рубина	13,07	13,04	13,10	87
					Усреднение	*13,07*	*12,85*	*13,29*	*26*
		Дисперсия	7,286	9,592	Правило Рубина	8,61	8,55	8,67	95
					Усреднение	*8,077*	*7,009*	*9,245*	*27*
30	1000	Среднее	12,89	13,36	Правило Рубина	13,06	13,03	13,09	87
					Усреднение	*13,05*	*12,87*	*13,25*	*28*
		Дисперсия	7,306	9,581	*Правило Рубина*	*0,08*	*8,02*	*8,14*	*95*
					Усреднение	6,150	5,327	6,920	204

Окончание табл. 2

Доля пропусков, %	Извлеченные выборки	Параметр	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			верхняя	нижняя					
30	10 000	Среднее	12,90	13,36	Правило Рубина	13,06	13,03	13,09	87
					Усреднение	*13,05*	*12,86*	*13,25*	*33*
		Дисперсия	7,286	9,580	*Правило Рубина*	*0,08*	*8,02*	*8,14*	*95*
					Усреднение	6,150	5,387	6,954	197
	50 000	Среднее	12,90	13,36	Правило Рубина	13,06	13,03	13,09	87
					Усреднение	*13,05*	*12,86*	*13,25*	*33*
		Дисперсия	7,286	9,592	*Правило Рубина*	*0,08*	*8,02*	*8,14*	*95*
					Усреднение	6,150	5,367	6,966	197
50	1000	Среднее	12,89	13,36	Правило Рубина	13,05	13,02	13,08	87
					Усреднение	*13,05*	*12,88*	*13,23*	*30*
		Дисперсия	7,306	9,581	*Правило Рубина*	*8,85*	*8,78*	*8,92*	*94*
					Усреднение	4,894	4,276	5,546	311
	10 000	Среднее	12,90	13,36	Правило Рубина	13,05	13,02	13,08	87
					Усреднение	*13,05*	*12,88*	*13,23*	*33*
		Дисперсия	7,286	9,580	*Правило Рубина*	*8,85*	*8,78*	*8,92*	*94*
					Усреднение	4,894	4,249	5,552	308
	50 000	Среднее	12,90	13,36	Правило Рубина	13,05	13,02	13,08	87
					Усреднение	*13,05*	*12,88*	*13,23*	*33*
		Дисперсия	7,286	9,592	*Правило Рубина*	*8,85*	*8,78*	*8,92*	*94*
					Усреднение	4,894	4,258	5,568	306

Таблица 3

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ПРАВИЛА РУБИНА И УСРЕДНЕНИЯ ПОДСТАВЛЕННЫХ ЗНАЧЕНИЙ ДЛЯ ОЦЕНКИ ПАРНЫХ КОЭФФИЦИЕНТОВ СВЯЗИ В СИТУАЦИЯХ НАЛИЧИЯ И ОТСУТСТВИЯ СВЯЗИ

Метод анализа данных	Связь	Доля пропусков, %	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ						
				нижняя	верхняя			нижняя	верхняя	
Коэффициент *V* Крамера	Присутствует	10	1000	0,056	0,215	Правило Рубина	0,08	0,07	0,1	77
						Усреднение	*0,094*	*0,043*	*0,182*	*20*
			10 000	0,060	0,216	Правило Рубина	0,08	0,07	0,1	78
						Усреднение	*0,094*	*0,044*	*0,186*	*24*
			50 000	0,061	0,216	Правило Рубина	0,08	0,07	0,1	78
						Усреднение	*0,094*	*0,044*	*0,186*	*24*
		30	1000	0,056	0,215	Правило Рубина	0,09	0,07	0,1	77
						Усреднение	*0,098*	*0,042*	*0,195*	*29*
			10 000	0,060	0,216	Правило Рубина	0,09	0,07	0,1	78
						Усреднение	*0,098*	*0,044*	*0,193*	*29*
			50 000	0,061	0,216	Правило Рубина	0,09	0,07	0,1	78
						Усреднение	*0,098*	*0,044*	*0,193*	*29*
		50	1000	0,056	0,215	*Правило Рубина*	*0,12*	*0,1*	*0,13*	*77*
						Усреднение	*0,141*	*0,08*	*0,224*	*79*
			10 000	0,060	0,216	*Правило Рубина*	*0,12*	*0,1*	*0,13*	*78*
						Усреднение	*0,141*	*0,079*	*0,226*	*79*
			50 000	0,061	0,216	*Правило Рубина*	*0,12*	*0,1*	*0,13*	*78*
						Усреднение	*0,141*	*0,08*	*0,228*	*81*

Продолжение табл. 3

Метод анализа данных	Связь	Доля пропусков, %	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Коэффициент *V* Крамера	Отсутствует	10	1000	0,033	0,166	Правило Рубина	–0,01	–0,02	0,01	81
						Усреднение	*0,026*	*–0,062*	*0,105*	*37*
			10 000	0,031	0,167	Правило Рубина	–0,01	–0,02	0,01	80
						Усреднение	*0,026*	*–0,054*	*0,105*	*41*
			50 000	0,031	0,167	Правило Рубина	–0,01	–0,02	0,01	81
						Усреднение	*0,026*	*–0,051*	*0,107*	*43*
		30	1000	0,033	0,166	*Правило Рубина*	*–0,06*	*–0,08*	*0,04*	*25*
						Усреднение	–0,091	–0,168	0,01	89
			10 000	0,031	0,167	*Правило Рубина*	*–0,06*	*–0,08*	*0,04*	*22*
						Усреднение	–0,091	–0,167	0,012	94
			50 000	0,031	0,167	*Правило Рубина*	*–0,06*	*–0,08*	*0,04*	*23*
						Усреднение	–0,091	–0,167	0,012	93
		50	1000	0,033	0,166	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,059*	*–0,138*	*0,016*	*66*
			10 000	0,031	0,167	Правило Рубина	–0,04	–0,05	–0,02	80
						Усреднение	*–0,059*	*–0,137*	*0,017*	*71*
			50 000	0,031	0,167	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,059*	*–0,137*	*0,017*	*70*

Продолжение табл. 3

Метод анализа данных	Связь	Доля пропусков, %	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Коэффициент корреляции Спирмена	Присутствует	10	1000	–0,458	–0,321	Правило Рубина	–0,33	–0,35	–0,31	87
						Усреднение	*–0,346*	*–0,423*	*–0,270*	*63*
			10 000	–0,460	–0,320	Правило Рубина	–0,33	–0,35	–0,31	86
						Усреднение	*–0,346*	*–0,421*	*–0,269*	*64*
			50 000	–0,460	–0,320	Правило Рубина	–0,33	–0,35	–0,31	86
						Усреднение	*–0,346*	*–0,419*	*–0,269*	*66*
		30	1000	–0,458	–0,321	*Правило Рубина*	*–0,16*	*–0,18*	*–0,14*	*335*
						Усреднение	–0,141	–0,224	–0,056	364
			10 000	–0,460	–0,320	*Правило Рубина*	*–0,16*	*–0,18*	*–0,14*	*329*
						Усреднение	–0,141	–0,224	–0,056	357
			50 000	–0,460	–0,320	*Правило Рубина*	*–0,16*	*–0,18*	*–0,14*	*329*
						Усреднение	–0,141	–0,224	–0,056	357
		50	1000	–0,458	–0,321	Правило Рубина	–0,04	–0,05	–0,02	518
						Усреднение	*–0,151*	*–0,229*	*–0,069*	*351*
			10 000	–0,460	–0,320	Правило Рубина	–0,04	–0,05	–0,02	507
						Усреднение	*–0,151*	*–0,228*	*–0,074*	*341*
			50 000	–0,460	–0,320	Правило Рубина	–0,04	–0,05	–0,02	507
						Усреднение	*–0,151*	*–0,228*	*–0,075*	*341*

Продолжение табл. 3

Метод анализа данных	Связь	Доля пропусков, %	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Коэффициент корреляции Спирмена	Отсутствует	10	1000	–0,088	0,072	Правило Рубина	–0,01	–0,02	0,01	81
						Усреднение	*0,026*	*–0,062*	*0,105*	*37*
			10 000	–0,082	0,071	Правило Рубина	–0,01	–0,02	0,01	80
						Усреднение	*0,026*	*–0,054*	*0,105*	*41*
			50 000	–0,083	0,072	Правило Рубина	–0,01	–0,02	0,01	81
						Усреднение	0,026	–0,051	0,107	43
		30	1000	–0,088	0,072	*Правило Рубина*	*–0,06*	*–0,08*	*0,04*	*25*
						Усреднение	–0,091	–0,168	0,01	89
			10 000	–0,082	0,071	*Правило Рубина*	*–0,06*	*–0,08*	*0,04*	*22*
						Усреднение	–0,091	–0,167	0,012	94
			50 000	–0,083	0,072	*Правило Рубина*	*–0,06*	*–0,08*	*0,04*	*23*
						Усреднение	–0,091	–0,167	0,012	93
		50	1000	–0,088	0,072	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,059*	*–0,138*	*0,016*	*66*
			10 000	–0,082	0,071	Правило Рубина	–0,04	–0,05	–0,02	80
						Усреднение	*–0,059*	*–0,137*	*0,017*	*71*
			50 000	–0,083	0,072	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,059*	*–0,137*	*0,017*	*70*

Продолжение табл. 3

Метод анализа данных	Связь	Доля пропусков, %	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Коэффициент корреляции Пирсона	Присутствует	10	1000	0,276	0,417	Правило Рубина	0,2	0,19	0,22	201
						Усреднение	*0,244*	*0,175*	*0,318*	*142*
			10 000	0,273	0,415	Правило Рубина	0,2	0,19	0,22	196
						Усреднение	*0,244*	*0,168*	*0,317*	*143*
			50 000	0,273	0,415	Правило Рубина	0,2	0,19	0,22	196
						Усреднение	*0,244*	*0,167*	*0,317*	*144*
		30	1000	0,276	0,417	Правило Рубина	0,12	0,1	0,13	328
						Усреднение	*0,166*	*0,090*	*0,239*	*258*
			10 000	0,273	0,415	Правило Рубина	0,12	0,1	0,13	323
						Усреднение	*0,166*	*0,092*	*0,239*	*251*
			50 000	0,273	0,415	Правило Рубина	0,12	0,1	0,13	323
						Усреднение	*0,166*	*0,092*	*0,239*	*251*
		50	1000	0,276	0,417	Правило Рубина	0,03	0,02	0,05	442
						Усреднение	*0,110*	*0,036*	*0,178*	*340*
			10 000	0,273	0,415	Правило Рубина	0,03	0,02	0,05	435
						Усреднение	*0,110*	*0,037*	*0,184*	*329*
			50 000	0,273	0,415	Правило Рубина	0,03	0,02	0,05	435
						Усреднение	*0,110*	*0,037*	*0,184*	*329*

Окончание табл. 3

Метод анализа данных	Связь	Доля пропусков, %	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
				Граница ДИ				нижняя	верхняя	
				нижняя	верхняя					
Коэффициент корреляции Пирсона	Отсутствует	10	1000	–0,160	–0,001	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,047*	*–0,128*	*–0,030*	*38*
			10 000	–0,157	0,005	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,047*	*–0,125*	*–0,032*	*43*
			50 000	–0,156	0,005	Правило Рубина	–0,04	–0,05	–0,02	81
						Усреднение	*–0,047*	*–0,125*	*–0,032*	*42*
		30	1000	–0,160	–0,001	Правило Рубина	–0,05	–0,07	–0,03	75
						Усреднение	*–0,084*	*–0,160*	*–0,005*	*3*
			10 000	–0,157	0,005	Правило Рубина	–0,05	–0,07	–0,03	75
						Усреднение	*–0,084*	*–0,160*	*–0,007*	*9*
			50 000	–0,156	0,005	Правило Рубина	–0,05	–0,07	–0,03	75
						Усреднение	*–0,084*	*–0,160*	*–0,007*	*10*
		50	1000	–0,160	–0,001	Правило Рубина	–0,05	–0,06	–0,03	81
						Усреднение	*–0,106*	*–0,181*	*–0,029*	*31*
			10 000	–0,157	0,005	Правило Рубина	–0,05	–0,06	–0,03	81
						Усреднение	*–0,106*	*–0,182*	*–0,029*	*36*
			50 000	–0,156	0,005	Правило Рубина	–0,05	–0,06	–0,03	81
						Усреднение	*–0,106*	*–0,182*	*–0,029*	*37*

Таблица 4

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ПРАВИЛА РУБИНА И УСРЕДНЕНИЯ ПОДСТАВЛЕННЫХ ЗНАЧЕНИЙ ДЛЯ ОЦЕНКИ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ КОЭФФИЦИЕНТОВ

Доля пропусков, %	Член уравнения	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			нижняя	верхняя					
10	Константа (зависимая переменная – «положение в обществе»)	1000	2,434	4,197	Правило Рубина	4,05	4,00	4,11	225
					Усреднение	*4,887*	*3,731*	*6,029*	*210*
		10 000	2,377	4,167	Правило Рубина	4,05	3,98	4,13	223
					Усреднение	*4,887*	*3,771*	*5,971*	*306*
		50 000	2,379	4,164	Правило Рубина	4,05	3,98	4,13	318
					Усреднение	*4,887*	*3,747*	*5,996*	*305*
	Незначимый регрессионный коэффициент («истинная» интервальная шкала, переменная «длительность очного образования»)	1000	–0,028	0,060	*Правило Рубина*	*0,03*	*0,02*	*0,04*	*85*
					Усреднение	–0,037	–0,095	0,022	107
		10 000	–0,028	0,059	*Правило Рубина*	*0,03*	*0,02*	*0,04*	*87*
					Усреднение	–0,037	–0,092	0,019	128
		50 000	–0,028	0,059	*Правило Рубина*	*0,03*	*0,02*	*0,04*	*109*
					Усреднение	–0,037	–0,092	0,020	123
	Незначимый регрессионный коэффициент (11-балльная шкала, переменная «удовлетворенность системой образования»)	1000	–0,034	0,072	*Правило Рубина*	*0,04*	*0,02*	*0,05*	*79*
					Усреднение	0,142	0,075	0,204	132
		10 000	–0,029	0,073	*Правило Рубина*	*0,04*	*0,02*	*0,05*	*77*
					Усреднение	0,142	0,076	0,207	156
		50 000	–0,030	0,073	*Правило Рубина*	*0,04*	*0,02*	*0,05*	*101*
					Усреднение	0,142	0,08	0,205	159

Продолжение табл. 4

Доля пропусков, %	Член уравнения	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			нижняя	верхняя					
10	Значимый регрессионный коэффициент (11-балльная шкала, переменная – ответ на вопрос «Насколько вы счастливы?»)	1000	0,205	0,332	*Правило Рубина*	*0,17*	*0,16*	*0,19*	*197*
					Усреднение	0,022	–0,051	0,092	363
		10 000	0,206	0,333	*Правило Рубина*	*0,17*	*0,16*	*0,19*	*198*
					Усреднение	0,022	–0,049	0,092	450
		50 000	0,206	0,333	*Правило Рубина*	*0,17*	*0,16*	*0,19*	*286*
					Усреднение	0,022	–0,051	0,093	452
	Значимый регрессионный коэффициент («истинная» интервальная шкала, переменная «возраст»)	1000	–0,021	–0,005	*Правило Рубина*	*–0,02*	*–0,02*	*–0,01*	*138*
					Усреднение	–0,007	–0,017	0,003	156
		10 000	–0,021	–0,005	Правило Рубина	–0,02	–0,02	–0,01	138
					Усреднение	*–0,007*	*–0,017*	*0,002*	*56*
		50 000	–0,021	–0,005	*Правило Рубина*	*–0,02*	*–0,02*	*–0,01*	*38*
					Усреднение	–0,007	–0,017	0,003	56
30	Константа (зависимая переменная – «положение в обществе»)	1000	2,434	4,197	Правило Рубина	5,76	5,54	5,8	313
					Усреднение	*6,153*	*4,802*	*7,551*	*271*
		10 000	2,377	4,167	*Правило Рубина*	*5,76*	*5,54*	*5,8*	*310*
					Усреднение	6,153	4,740	7,592	360
		50 000	2,379	4,164	Правило Рубина	5,76	5,54	5,8	405
					Усреднение	*6,153*	*4,729*	*7,575*	*360*

Продолжение табл. 4

Доля пропусков, %	Член уравнения	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			нижняя	верхняя					
30	Незначимый регрессионный коэффициент («истинная» интервальная шкала, переменная «длительность очного образования»)	1000	–0,028	0,060	*Правило Рубина*	*0*	*–0,01*	*0,02*	*51*
					Усреднение	–0,081	–0,158	–0,007	178
		10 000	–0,028	0,059	*Правило Рубина*	*0*	*–0,01*	*0,02*	*53*
					Усреднение	–0,081	–0,154	–0,008	199
		50 000	–0,028	0,059	*Правило Рубина*	*0*	*–0,01*	*0,02*	*75*
					Усреднение	–0,081	–0,154	–0,008	194
	Незначимый регрессионный коэффициент (11-балльная шкала, переменная «удовлетворенность системой образования»)	1000	–0,034	0,072	*Правило Рубина*	*–0,03*	*–0,04*	*–0,01*	*34*
					Усреднение	0,057	–0,009	0,123	53
		10 000	–0,029	0,073	*Правило Рубина*	*–0,03*	*–0,04*	*–0,01*	*40*
					Усреднение	0,057	–0,009	0,123	73
		50 000	–0,030	0,073	*Правило Рубина*	*–0,03*	*–0,04*	*–0,01*	*62*
					Усреднение	0,057	–0,009	0,124	73
	Значимый регрессионный коэффициент (11-балльная шкала, переменная – ответ на вопрос «Насколько вы счастливы?»)	1000	0,205	0,332	*Правило Рубина*	*0,11*	*0,09*	*0,12*	*252*
					Усреднение	0,007	–0,069	0,083	377
		10 000	0,206	0,333	*Правило Рубина*	*0,11*	*0,09*	*0,12*	*254*
					Усреднение	0,007	–0,066	0,081	464
		50 000	0,206	0,333	*Правило Рубина*	*0,11*	*0,09*	*0,12*	*341*
					Усреднение	0,007	–0,069	0,081	466

Продолжение табл. 4

Доля пропусков, %	Член уравнения	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			нижняя	верхняя					
30	Значимый регрессионный коэффициент («истинная» интервальная шкала, переменная «возраст»)	1000	–0,021	–0,005	Правило Рубина	–0,02	–0,03	–0,01	188
					Усреднение	*–0,07*	*–0,028*	*0,004*	*175*
		10 000	–0,021	–0,005	Правило Рубина	–0,02	–0,03	–0,01	188
					Усреднение	*–0,07*	*–0,028*	*0,004*	*75*
		50 000	–0,021	–0,005	Правило Рубина	–0,02	–0,03	–0,01	88
					Усреднение	*–0,07*	*–0,019*	*0,004*	*44*
50	Константа (зависимая переменная – «положение в обществе»)	1000	2,434	4,197	*Правило Рубина*	*4,60*	*4,50*	*4,71*	*254*
					Усреднение	6,754	5,248	8,180	296
		10 000	2,377	4,167	*Правило Рубина*	*4,60*	*4,50*	*4,71*	*252*
					Усреднение	6,754	5,386	8,142	396
		50 000	2,379	4,164	*Правило Рубина*	*4,60*	*4,50*	*4,71*	*347*
					Усреднение	6,754	5,367	8,166	396
	Незначимый регрессионный коэффициент («истинная» интервальная шкала, переменная «длительность очного образования»)	1000	–0,028	0,060	*Правило Рубина*	*0,02*	*0,00*	*0,04*	*63*
					Усреднение	–0,112	–0,191	–0,032	216
		10 000	–0,028	0,059	*Правило Рубина*	*0,02*	*0,00*	*0,04*	*64*
					Усреднение	–0,112	–0,191	–0,035	241
		50 000	–0,028	0,059	*Правило Рубина*	*0,02*	*0,00*	*0,04*	*86*
					Усреднение	–0,112	–0,190	–0,036	236

Окончание табл. 4

Доля пропусков, %	Член уравнения	Извлеченные выборки	Эталон		Способ агрегирования	Точечная оценка	Граница ДИ		Δ, %
			Граница ДИ				нижняя	верхняя	
			нижняя	верхняя					
50	Незначимый регрессионный коэффициент (11-балльная шкала, переменная «удовлетворенность системой образования»)	1000	–0,034	0,072	*Правило Рубина*	*–0,04*	*–0,06*	*–0,03*	*53*
					Усреднение	*–0,008*	*–0,063*	*0,051*	*57*
		10 000	–0,029	0,073	*Правило Рубина*	*–0,04*	*–0,06*	*–0,03*	*60*
					Усреднение	–0,008	–0,067	0,050	90
		50 000	–0,030	0,073	*Правило Рубина*	*–0,04*	*–0,06*	*–0,03*	*82*
					Усреднение	–0,008	–0,065	0,049	86
	Значимый регрессионный коэффициент (11-балльная шкала, переменная – ответ на вопрос «Насколько вы счастливы?»)	1000	0,205	0,332	*Правило Рубина*	*0,04*	*0,03*	*0,06*	*299*
					Усреднение	–0,011	–0,099	0,077	401
		10 000	0,206	0,333	*Правило Рубина*	*0,04*	*0,03*	*0,06*	*301*
					Усреднение	–0,011	–0,102	0,079	492
		50 000	0,206	0,333	*Правило Рубина*	*0,04*	*0,03*	*0,06*	*388*
					Усреднение	–0,011	–0,098	0,080	489
	Значимый регрессионный коэффициент («истинная» интервальная шкала, переменная «возраст»)	1000	–0,021	–0,005	Правило Рубина	0,00	0,00	0,01	263
					Усреднение	*–0,006*	*–0,018*	*0,006*	*150*
		10 000	–0,021	–0,005	Правило Рубина	0,00	0,00	0,01	263
					Усреднение	*–0,006*	*–0,019*	*0,006*	*44*
		50 000	–0,021	–0,005	Правило Рубина	0,00	0,00	0,01	163
					Усреднение	*–0,006*	*–0,019*	*0,006*	*44*

Zangieva Irina,
National Research University Higher School of Economics (NRU HSE), Moscow, izangieva@hse.ru

Suleymanova Anna,
National Research University Higher School of Economics (NRU HSE), Moscow, ansuleymanova@edu.hse.ru

Comparative analysis of approaches to aggregation of multiple imputation results

Multiple imputation is an approach to missing data elimination created by Donald Rubin. The purpose of multiple imputation is to reconstruct the initial structure of data, i.e. to generate the answers as close as possible to hypothetical complete dataset. However, the original algorithm of multiple imputation is complicated and demands a major amount of effort to accomplish. In the study simpler alternative approach –averaging of imputed values – was experimentally tested against Rubin's rule in a number of common research situations. We compared two approaches to multiple imputation results aggregation – Rubin's rule and averaging of imputed values – considering given analytical tools, share of missing values and type of the variable that contains missing values. The results were summed up in a set of recommendations describing a pertinent approach to aggregation for each research situation.

Keywords: missing data, item nonresponse, multiple imputation, Rubin's rule, averaging of imputed values, research situation

References

1. Zangieva I.K. "Problema propuskov v sotsiologicheskih dannyh: smysl i podhody k resheniyu" (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2012, 33, 28–56.
2. Brand J.P.L. *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. Thesis Erasmus University Rotterdam, 1999.
3. Jin H., Rubin D. "Public Schools versus Private Schools: Causal Inference with Partial Compliance", *Journal of Educational and Behavioral Statistics*, 2009, 34 (1), 24– 45.

4. Zangieva I.K., Tolstova Yu.N. "Ponyatie sluchajnosti i problema propuskov dannyh v sociologii" (in Russian), in: *Matematicheskoe modelirovanie social'nyh processov*. M.: Sociologicheskij fakul'tet MGU, 2012. Vyp. 14. Gl. 14. P. 146–165.
5. Rubin D. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, 1987.
6. Rubin D. "The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys", *The American Statistician*, 2004, 58 (4), 298–302.
7. Mitra R., Reiter J.P. "A comparison of two methods of estimating propensity scores after multiple imputation", *Statistical Methods in Medical Research*, 2016, 25 (1), 188–204.
8. Vink G., van Buuren S. "Pooling multiple imputations when the sample happens to be the population" [online source], in: *Cornell University Library*. 2014. URL: http://arxiv.org/abs/1409.8542 (date of access: May 3, 2016).
9. Zhang P. "Multiple imputation: theory and method", *International Statistical Review*, 2003, 71 (3), 581–592.
10. Kutlaliev A. H. "Metod mnozhestvennogo vosstanovleniya dannyh" (in Russian), in: *Sociologicheskie metody v sovremennoj issledovatel'skoj praktike*. M.: HRU HSE, 2011. P. 201–208.
11. Kromrey J.D., Hines C.V. "Nonrandomly Missing Data in Multiple Regression: An Empirical Comparison of Common Missing-Data Treatments", *Educational and Psychological Measurement*, 2003, 54, 573–593.
12. Shitikov V.K., Rozenberg G.S. *Randomizaciya i butstrep: statisticheskij analiz v biologii i ekologii s ispol'zovaniem R* (in Russian). Tol'yatti: Kassandra, 2013.
13. Efron B. "Bayesian inference and the parametric bootstrap", *The Annals of Applied Statistics*, 2012, 6 (4), 1971–1997.
14. Jacoby W., Armstrong D. "II Bootstrap Confidence Regions for Multidimensional Scaling Solutions", *American Journal of Political Science*, 2014, 58 (1), 264–278.
15. Manly B. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Third Edition. Chapman and Hall / CRC, 2006.
16. Zangieva I.K., Timonina E.S. "Sravnenie effektivnosti algoritmov zapolneniya propuskov v dannyh v zavisimosti ot ispol'zuemogo metoda analiza" (in Russian), *Monitoring obshhestvennogo mneniya: jekonomicheskie i social'nye peremeny (The monitoring of public opinion: economic and social changes journal)*, 2014, 1 (119), 41–55.