*К.Ю. Терентьев  
ст. преп. каф. социологии,   
зав. лабораторией социологических исследований  
Петрозаводского государственного университета*

**Создание базы данных анкетного опроса:  
возможности MS Excel**

Первый шаг обработки данных анкетного опроса - формирование электронной (компьютерной) базы данных. Для этих целей используются либо специализированные программы для обработки статистических данных, такие как SPSS или Statistica, или же более универсальные программы, например MS Excel или Access (или аналогичные программы, например, из пакета Open Office).

Очевидно, что узкоспециализированные программы имеют целый ряд преимуществ, поскольку «заточены» под решение определенных задач, в них нет ничего «лишнего» - соответственно, при прочих равных, предпочтительнее использовать именно эти программы для работы с социологическими (статистическими) данными. Но в ряде случаев и универсальные программы оказываются весьма полезными. Преимущество универсальных программ – их распространенность, присутствие на большинстве компьютеров, а также собственно универсальность – возможность решать разнообразные задачи. В данной работе представлены размышления об использовании программы MS Excel как средства формирования баз данных количественных социологических исследований, основанные на опыте работы лаборатории социологических исследований Петрозаводского государственного университета (ПетрГУ)[[1]](#footnote-1).

На сегодняшний день сотрудники лаборатории активно используют программу SPSS для работы с данными количественных социологических исследований, однако это не привело к отказу от программы MS Excel. Среди основных направлений использования можно выделить:

* **Формирование баз данных, ввод данных.** Базы, сформированные в MS Excel, довольно легко импортируются в SPSS. Некоторые проблемы могут возникнуть из-за открытых вопросов, а также заголовков столбцов, которые в SPSS не нужны, точнее фиксируются по другому. Однако эти проблемы совершенно не критичны и легко преодолеваются. А в ситуации, когда большинство компьютеров не оснащены SPSS, формирование базы в MS Excel значительно ускоряет ввод данных;
* **Работа с открытыми вопросами.** Обработка, перекодировка и редактирование текстовых данных в Excel значительно проще, чем, например, в SPSS;
* **Создание графиков и диаграмм.** Создание, простота работы и наглядность, на наш взгляд, в MS Excel реализованы лучше, чем в SPSS. Особенно в версиях 2007 и 2010;
* **Редактирование и форматирование таблиц для импорта в текстовый редактор.** Зачастую бывает необходимо переформатировать таблицы, подготовленные в SPSS (например, убрать какие-то ячейки, разделить или объединить их). Это значительно проще сделать в MS Excel, чем в самом SPSS или, например, в MS Word.

Вышеназванное обуславливает сегодняшнюю актуальность заявленных размышлений.

В конце 2007 года лабораторией социологических исследований ПетрГУ (руководитель К.Ю. Терентьев) по заказу Министерства Республики Карелия по вопросам национальной политики и связям с религиозными объединениями был разработан инструментарий проведения социологического мониторинга «Динамика изменений социальных установок в общественном сознании населения Республики Карелия как результат реализации региональной целевой программы «Гармонизация национальных и конфессиональных отношений, формирование гражданского согласия в Республике Карелия на 2007-2011 годы» («Карелия – территория согласия»)». Одним из требований заказчика было описание процедуры внесения данных собранных анкет в электронную базу данных. В одном из приложений к программе исследования это описание было представлено (*Приложение 2. Основные правила внесения данных анкеты в сводную матрицу базы данных анкетирования*). В силу того, что на тот момент программа SPSS не использовалась в ПетрГУ (равно как и в министерстве), а основной программой для обработки баз данных была MS Excel, то описание было сделано именно под программное обеспечение корпорации Microsoft.

Правила, сформулированные в этом Приложении, активно применяются лабораторией и сейчас, конечно с некоторыми поправками, подсказанными позднейшей практикой обработки данных социологических исследований в MS Excel и SPSS. Ниже приведен перечень этих правил или, что даже точнее, алгоритм составления матрицы базы данных в MS Excel.

В большинстве социологических исследований анализируется анкетная информация. Вообще, «в отечественной социологии самым популярным является индивидуальный анкетный опрос. Им пользуются приблизительно в 75—80% случаев»[[2]](#footnote-2). Полученную с помощью этого метода информацию при переносе из бумажного вида в электронную базу данных можно представить в виде матрицы, строкам которой соответствуют объекты, или наблюдения (респонденты, анкеты), а столбцам - переменные (отдельные вопросы анкеты или варианты ответов).

Надо сказать, что в ходе работы с данными социологических исследований, нам встречались также и попытки создать базы, организованные по обратному принципу: в строках – переменные, в столбцах – случаи. Однако такой вариант формирования базы представляется неприемлемым, поскольку:

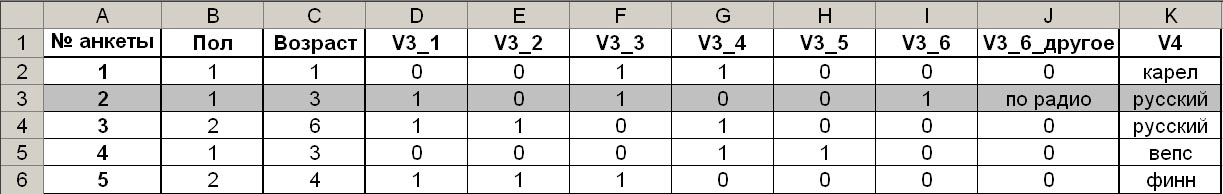
* на одном листе MS Excel только 256 столбцов, а это совершенно несопоставимо с размерами выборочных совокупностей большинства анкетных социологических исследований (в версии MS Excel 2010 этого ограничения уже нет)
* подсчитывать и анализировать данные по строкам в MS Excel довольно трудоемко, а в отдельных случаях невозможно;
* к тому же будет попросту невозможен прямой импорт данных в специализированные программы обработки данных, такие как SPSS



***Рис. 1 Пример вопросов анкеты[[3]](#footnote-3).***

При кодировании информации удобно пользоваться определенными правилами заполнения матрицы в соответствии со структурой обрабатываемой анкеты.

Матрица данных, собранных на основании такой анкеты, изображена на рис.2. Пол здесь закодирован в соответствии с содержимым анкеты кодами «1» - мужчины, «2» - женщины; возраст цифрами от 1 до 6 – номерами переменных (ответов), которые указал респондент; ответ на третий вопрос кодируется по каждой переменной (варианту ответа) двоичной системой кодировки (дихотомией): «1» – если вариант выбран и «0»[[4]](#footnote-4) - если нет; национальность – вписывается тот вариант, который респондент записал сам.

****

***Рис. 2. Структура матрицы базы данных опроса общественного мнения населения***

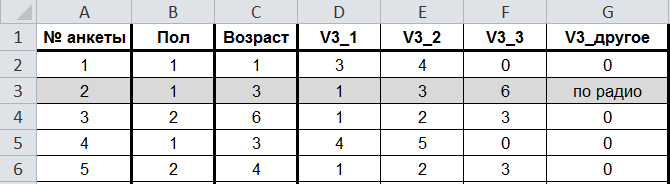
Основные правила подготовки базы данных:

1. Каждая анкета перед внесением данных в базу пронумеровывается порядковым номером (сквозная нумерация для всего массива анкет). Соответственно база данных всегда должна содержать столбец (как правило, первый), в котором фиксируется номер вводимых анкет. На рис. 2 это столбец A:A «№ анкеты». Необходимость фиксации номера вводимой анкеты очевидна – это позволяет в любой момент обратиться к соответствующей бумажной анкете и сверить данные. Использование же для этой цели программной нумерации строк невозможно, в силу, во-первых, смещения нумерации из-за строки заголовков (соответственно строка 3 листа MS Excel будет содержать данные анкеты №2) и, во-вторых, процедура отбраковки анкет может привести к изъятию из общего массива уже пронумерованных анкет, что собьет сквозную нумерацию.
2. Данные одной анкеты вводятся по строке (выделенная строка в таблице на рис. 2 – это данные одного респондента – одной анкеты)
3. Заголовки столбцов должны отличаться от закодированных числовых данных и идентифицировать переменную. Можно рекомендовать или называть их отдельным словом (словосочетанием), или, если используется номер вопроса, использовать перед ним буквенный символ (например, кириллическую «В», латинские «V» или «Q»). Не рекомендуется использовать в заголовках столбцов пробелы. Всё это необходимо, чтобы программа понимала их именно как заголовки столбцов, а также так удобнее работать со списками переменных при формировании сводных таблиц.
4. Вопросы, на которые респондент может выбрать только один из предложенных вариантов ответов, кодируются в одном столбце номером выбранного ответа – например, пол или возраст (см. рис. 1 и 2). Если респондент мужчина – в столбец **«Пол»** ставится «1», если женщина – «2». Если респондент вообще не выбрал ни одного варианта ответа, в ячейку ставится «0». Также кодируется и возраст респондента. Соответственно, из таблицы видно, что респондент №2 (его ответы в таблице выделены цветом) – мужчина в возрасте между 36 и 45 годами.
5. Для вопросов, на которые респондент может выбрать одновременно несколько вариантов ответа, в базе данных создается такое количество столбцов, сколько всего вариантов ответа предлагается на этот вопрос в анкете. Соответственно, для вопроса №3 из примера анкеты (рис.1) в матрице необходимо создать 6 столбцов (рис. 2)[[5]](#footnote-5). Данные анкеты кодируются проставлением цифры «1» в ячейке столбца, соответствующего выбранному варианту, если этот вариант ответа на вопрос выбран респондентом, и цифрой «0» - если вариант не был выбран.

Например, на вопрос №3 об оскорбительных анекдотах респондент № 2 ответил (как можно видеть из базы данных на рис. 2), что подобные анекдоты он слышал в компании знакомых, а также читал в газетах и журналах. Соответственно, его ответы заносятся в базу так: в ячейках столбцов, соответствующих выбранным ответам (V3\_1 (D3) и V3\_3 (F3)) отмечаем значение «1», в остальных – «0». Также он отметил в варианте «Другое», что слышал анекдоты по радио (см. п.7).

Вопросы с множественными ответами, безусловно, можно кодировать не только с помощью дихотомии, но и через категории, как это происходит в SPSS. Особенно это уместно в вопросах со значительным количеством вариантов ответов, выбор которых для респондента ограничен неким числом (например, когда из длинного перечня общественных проблем респонденту предлагается назвать три-пять, по его мнению, наиболее существенных). В таком случае ответы на вопрос вносятся в пять (или шесть, на случай если ряд респондентов все-таки назовут большее количество вариантов ответа) столбцов с указанием номера выбранного ответа в каждом отдельном столбце (см. рис 3).

Однако, несмотря на все удобства такого варианта для ввода, обрабатывать данные, сформированные таким образом, в MS Excel довольно трудоемко, тогда как дихотомии легко подсчитываются с помощью сводных таблиц. Соответственно вариант с категориями возможен только в случае импорта данных из MS Excel в SPPS (например, когда вносить данные необходимо быстро, и для этого используются компьютеры без SPSS).



***Рис. 3 Внесение данных вопроса с множественным выбором через категории***

1. В табличных вопросах (см. рис. 4), когда необходимо дать оценку разных параметров по одному вопросу, в базу данных вносится выбранный респондентом номер ответа для каждого варианта в свой столбец. Иными словами, каждый вариант ответа табличного вопроса кодируется также, как если бы это был отдельный вопрос. Более корректно было бы указать, что табличные вопросы логично разделяются на «подвопросы». В примере на рис. 4, табличный вопрос о конфликтах разбивается на ряд «подвопросов»: «Как Вы оцениваете конфликтные отношения между богатыми и бедными?», «Как Вы оцениваете конфликтные отношения между Москвой и регионами?» и т.д. В свою очередь, ответы на них вносятся в соответствии с п.4 (если можно выбрать только один вариант ответа) или п.5 (если можно выбрать несколько вариантов ответа) – хотя второй вариант встречается довольно редко.



***Рис. 4 Пример вопросов анкеты (продолжение).***

В вопросе о конфликтах – табличном – необходимо оценить напряженность различных общественных конфликтов. Респондент №2 (в таблице на рис. 5 - выделен цветом) конфликт «между богатыми и бедными» оценил как острый (в примере на рис. 4 – это выделенный цветом вариант ответа), поэтому в соответствующую ячейку столбца **«V5\_1»** вносим цифру «2». Соответственно, в ячейку **«V5\_2»** вносим цифру «3», и т.д. Если респондент затруднился дать ответ на поставленный вопрос (выбрал вариант «трудно сказать» или «затрудняюсь ответить» в табличном вопросе), то в ячейку ставим «9» - как в **«V5\_3».** Если же респондент вообще не выбрал ни одного варианта ответа – ставим «0» - как в **«V5\_4».**

******

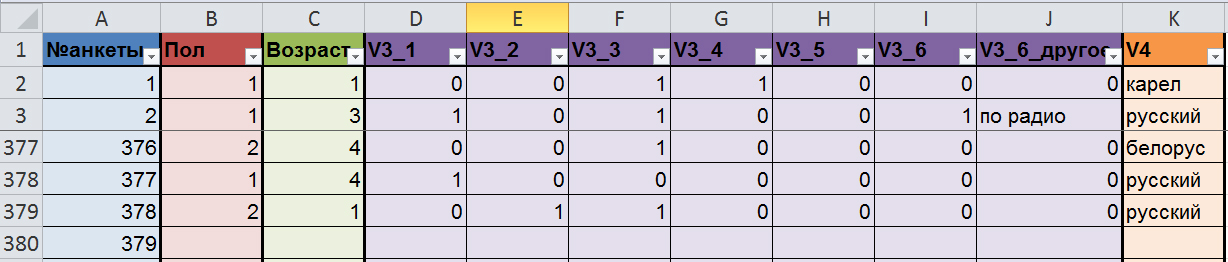
***Рис.5 Внесение в базу данных ответов на табличные вопросы***

1. Открытые вопросы (вопросы, где респондент сам «от руки» вписывает ответ), или вариант «Другое» в полузакрытых вопросах на этапе ввода данных не кодируются и записываются так, как это сделал сам респондент, дословно, вплоть до грамматических и орфографических ошибок. В отдельных случаях (в первую очередь, в полузакрытых вопросах с вариантом «Другое») уместно использовать дополнительный столбец для кодирования такой переменной (фиксации факта выбора этого варианта ответа): ставим «1» - если вариант «Другое» выбран, и «0» - если не выбран. От дополнительного столбца на этапе ввода, конечно, можно отказаться. Тогда его имеет смысл создать на этапе подготовки базы данных к обработке, поскольку MS Excel эффективнее подсчитывает числовые значения (не говоря уже о смысловом, а не только количественном кодировании открытого вопроса).

В примере на рис. 1 это вопросы об анекдотах и о национальности. В столбце «**V3\_6»** – ставим единицу в строке второй анкеты, поскольку в этой анкете респондент выбрал вариант другое, в следующем столбце **«V3\_6\_другое»** вписываем вариант, предложенный респондентом – «по радио». В столбце «**V4**» необходимо вписать ту национальность, которую указал респондент – карел, русский и т.д. (см. рис. 2).

1. Также можно сформулировать ряд необязательных правил, применение которых облегчает работу по вводу данных в базу.

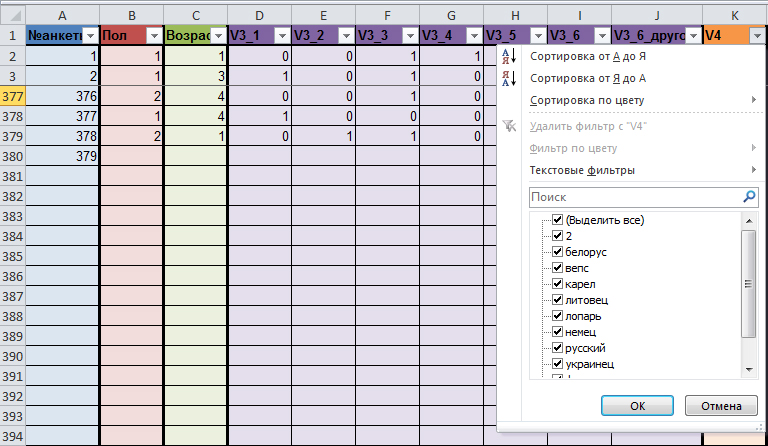
* Для удобства ввода данных (особенно при вводе анкет с большим количеством вопросов с множественными ответами), уместно графически отделить друг от друга столбцы отдельных вопросов. Это можно сделать заливкой разного цвета или более широкими границами столбцов (или даже использовать и то, и другое). Тогда при вводе разные вопросы легче отличать друг от друга (см. рис. 6).



***Рис. 6 Оформление матрицы базы данных в MS Excel***

* Также для удобства ввода данных большого числа анкет стоит использовать возможность закрепления строки, что позволит оставлять на виду заголовки столбцов вне зависимости от номера вводимой анкеты (на рис. 6 закреплена третья строка).
* При значительном числе вопросов с множественными ответами уместно при начале ввода данных путем копирования заполнить все ячейки нулями. Тогда вносить необходимо будет только выбранные ответы.

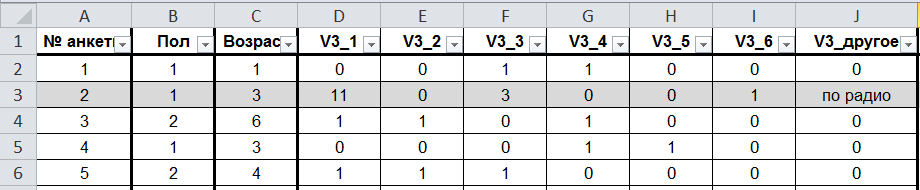
1. Для контроля ввода данных и их последующей перекодировки можно использовать функцию «Автофильтр», которая не только показывает все варианты переменной, внесенные в столбец (это, например, позволяет отсеять ряд «ошибочных» вариантов, попавшие в столбец при вводе), но и позволяет задавать условия отбора. (На рис. 7 автофильтр применен к данным вопроса 4. Очевидно, например, что в одной (или нескольких) из ячеек ошибочно поставлено значение «2», которое необходимо выделить в автофильтре, соотнести с анкетой и исправить).



***Рис. 7 Функция автофильтра (вопрос 4 (V4))***

Существует две наиболее типичные ошибки ввода, которые можно выявить с помощью функции «автофильтра». Во-первых, это ошибки, возникающие вследствие так называемого «залипания» клавиш, когда вместо одной в ячейке появляется две одинаковых цифры: вместо варианта «2» появляется «22», а вместо «5» - «55» (На рис. 8 в ячейке D3 вместо «1» внесено «11»). И если в номинальных переменных это будет выявлено и на этапе анализа (ошибки проявятся в качестве отдельных значений), то в случае анализа отдельных шкал (например, при подсчете среднего), такие ошибки могут серьезно исказить результат.

Во-вторых, это ошибки, связанные со случайным искажением принципа внесения данных вопросов с множественным выбором, когда вместо дихотомии в ячейки вносятся номера выбранных вариантов ответа (на рис. 8 в ячейку F3 вместо «1» внесено «3», поскольку в анкете респондент выбрал третий вариант ответа на вопрос №3). Если подобные ошибки не обнаружить и не исправить на этапе подготовки базы данных к обработке, то при подсчете это не будет замечено (ни при использовании сводных таблиц в MS Excel, ни при использовании функции *Multiple response* в SPSS), что приведет к искажению результатов.



***Рис. 8 Типичные ошибки ввода***

Использование вышеописанных правил позволяет оптимально использовать возможности программы MS Excel на этапе подготовки базы данных количественного социологического исследования, ускорить формирование базы данных, используя одновременно труд нескольких операторов ввода на разных компьютерах без специализированных статистических программ, существенным недостатком которых является их высокая стоимость. Навыки эффективной работы в универсальном табличном редакторе позволят значительно расширить возможности социолога по обработке, анализу и представлению статистических данных.

**Литература**

1. *Добреньков В.И., Кравченко А.И.* Методы социологического исследования: Учебник. — М.: ИНФРА-М, 2004. — 768 с.

1. Лаборатория социологических исследований факультета политических и социальных наук ПетрГУ создана в начале 2006 года, в первую очередь, для оценки качества образования в вузе. Также активно сотрудничает в проведении социологических исследований с органами государственной и муниципальной власти Республики Карелия. [↑](#footnote-ref-1)
2. Добреньков В.И., Кравченко А.И. Методы социологического исследования: Учебник. — М.: ИНФРА-М, 2004. С. 77 [↑](#footnote-ref-2)
3. Вопросы для примеров взяты из анкеты «Особенности межнациональных и межрелигиозных отношений в РК» вышеупомянутого исследования [↑](#footnote-ref-3)
4. В большинстве случаев, нулем можно обозначать в базе данных отсутствие ответа на вопрос (или отсутствие выбора конкретного варианта ответа). Однако иногда «0» может быть использован при кодировке других значений, а не только при отсутствии ответа на вопрос. Например, при вопросе о доходах 0 может означать отсутствие дохода, а в шкале оценки удовлетворенности, например, от -2 до +2, 0 – это нейтральное отношение. В таких случаях, отсутствие ответа рекомендуется кодировать специальным значением (например, 999997), либо попросту оставлять пустой соответствующую ячейку. [↑](#footnote-ref-4)
5. Также можно создавать дополнительный столбец для открытого варианта ответа («Другое») для одновременной фиксации факта выбора этого варианта, а также информации (текстовых данных) от респондента. Подробнее см. п.7 [↑](#footnote-ref-5)