Мартышенко Н.С., Мартышенко С.Н., Кустов Д.А. Инструментальные средства обработки анкетных данных в EXCEL // Информационные технологии моделирования и управления. – 2007. – №1. С. 112–120. 2007

ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ОБРАБОТКИ АНКЕТНЫХ ДАННЫХ В EXCEL

С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов
Владивостокский государственный университет экономики и сервиса
E-mail:kustov@vvsu.ru

В последнее время в нашей стране все большее распространение имеют исследования, основанные на анкетных опросах населения. Это связано, в первую очередь, с развитием маркетинговой деятельности предприятий. Однако, в разработке инструментальных средств обработки анкетных данных наблюдается значительное отставание от потребностей практики. Поэтому большинство исследователей вынуждены ограничиваться первичной обработкой данных, используя доступные инструментальные средства. Чаще всего используются средства обработки числовых данных EXCEL, реже – средства статистических пакетов, таких как STATISTICA или SPSS.

Данные анкетных опросов обладают рядом специфических особенностей. Поэтому для их обработки не всегда пригодны стандартные программные средства.

Специфика данных заключается в том, что они содержат большое количество нечисловой информации, порождаемой использованием в анкетах разнообразных измерительных шкал [3]. Большинство распространенных пакетов прикладных программ, напротив, нацелено на обработку числовой информации.

Наличие такого количества шкал вызвано стремлением исследователей получить от респондентов более достоверную информацию. Поскольку исследователь заинтересован в получении информации, ему и приходится подстраиваться под респондента, предоставляя вопросы в такой форме, в какой респондент сможет или пожелает ответить.

Как показали многочисленные опыты, человек более точно (и с меньшими затруднениями) отвечает на вопросы качественного, например, сравнительного, характера, чем количественного. Поэтому именно в анкетных опросов получили наибольшее распространение признаки нечисловой природы [2].

Последние годы мы активно занимались разработкой методов и инструментальных средств обработки анкетных данных. В настоящее время нами разработан специализированный комплекс программных средств, который позволяет решить ряд задач по обработке анкетных данных [1].

В данной работе мы предлагаем к рассмотрению средства решения одной из таких задач, а именно средства обработки данных открытых или неструктурированных вопросов. Такие данные являются наиболее сложными с точки зрения компьютерной обработки и в настоящее время практически отсутствуют инструментальные средства их обработки.

В отличие от закрытого вопроса, он не содержит подсказок, не “навязывает” тот или иной вариант ответа и рассчитан на получение неформализованного мнения. Еще чаще чем открытый вопрос, встречается полузакрытый вопрос, который кроме определенного числа вариантов ответа, содержит позицию “другое – укажите какое (что, где, как)”. Известны и иные формы открытого вопроса: “завершение предложения”, ”подбор ассоциации” и т.д.

Большинство исследователей не применяют компьютерную обработку данных открытых вопросов, а используют их в поисковых целях для получения информации для будущих исследований. Между тем, ответы на эти вопросы могут оказаться очень информативными.

При открытой форме вопроса можно было бы ожидать, что респонденты не дадут одинаковых ответов. На практике, перечень действительно различных по сути, а не по форме ответов на такие вопросы анкет ограничен. Уже при обработке порядка 700 анкет можно выделить всего 30–40 различных вариантов ответов. При увеличении объема выборки картина практически не изменяется. Выделенные варианты ответов можно интерпретировать как значения признака, измеренного в номинальной шкале.

Наличие 30–40 вариантов значений признака тоже слишком большое количество для анализа измерений в номинальной шкале. Поэтому исследователь после формирования приемлемого списка действительно различных вариантов ответов должен сгруппировать эти ответы, рассматривая их как некоторые характеристики непересекающихся классов (типов) респондентов.

Конечно, такое объединение будет носить субъективный характер, но, тем не менее, оно совершенно оправдано с точки зрения социологической теории личности, которая выделяет определенное количество типов личности. Это подтверждается большим количеством независимых исследований ученых из различных стран, которые приходили не более чем к 7-8 типам. В реальных исследованиях каждому из выделенных типов присваивается определенное название, ассоциированное с темой исследования. С математической точки зрения название не имеет никакого значения, а имеет смысл только операция объединения ряда значений признака в один класс. Поэтому типы (классы) могли бы быть просто пронумерованы в произвольном порядке.

Таким образом, с содержательной точки зрения операция преобразования открытого вопроса к номинальной шкале или иначе операция типизации не так уж и сложна. Однако, при переходе к практическим исследованиям с выборкой более 1000 респондентов и анкетами, содержащими более 30 вопросов, она становится достаточно трудоемкой.

Для решения этой задачи нами разработано специальное инструментальное средство, которое позволяет автоматизировать деятельность исследователя при поиске типологий по большим спискам первичных отве-

тов на открытый вопрос. Данное программное средство входит в состав разработанного нами специализированного комплекса программных средств обработки анкетных данных, предназначенного для работы в среде EXCEL [1], которое может быть использовано как самостоятельное средство. Подход, состоящий не в разработке собственного автономного пакета программных средств, а в расширении функций распространенного среди широкого круга практиков пакета, на наш взгляд, наиболее отвечает сегодняшнему уровню использования программных средств по обработке данных. Даже разрабатывая собственную технологию решения специфических задач по обработке анкетных данных, мы можем использовать всю мощь пакета EXCEL, как для выполнения отдельных промежуточных операций, так и для оформления результатов.

Разработанный программный модуль позволяет решать не только задачу типизации в простейшем случае, которая была рассмотрена выше, но и допускает решение более сложных задач, встречающихся на практике. Некоторые, более общие варианты постановок задач типизации, мы рассмотрим ниже.

Вначале рассмотрим работу программного модуля при решении простой задачи типизации. Поскольку программный модуль предназначен для работы в среде EXCEL, то его принципы работы и возможности должны быть продемонстрированы в этой программной среде.

Решить задачу типизации значений признака, порожденного открытым вопросом, можно и стандартными средствами EXCEL, используя для этого функции сортировки и корректировки данных. Однако при больших объемах выборки такой способ будет весьма трудоемким. Один и тот же ответ можно выразить десятками способов. Даже различие в одном символе EXCEL воспринимает, как разные значения. Достаточно поменять порядок слов и один и тот же ответ окажется в различных частях отсортированного списка.

В разработанном нами программном модуле мы попытались учесть все особенности задачи типизации, что позволило на порядок сократить время получения конечного результата по сравнению с решением задачи стандартными средствами EXCEL.

Работа с программным модулем начинается с отбора признака подлежащего типизации (рис. 1). При использовании программы в составе специализированного программного комплекса в списке признаков будут указаны только признаки, соответствующие открытым вопросам.

Программа формирует на отдельном листе EXCEL рабочую таблицу типизации, включающую четыре столбца. В первом столбце содержится список неповторяющихся значений признака (уникальных значений), второй – отведен для ввода названий классов, третий – для ввода названий подклассов, а в четвертом – выводятся частоты повторяемости уникальных

ответов. В исходном состоянии второй и третий столбцы не заполнены (рис. 2).

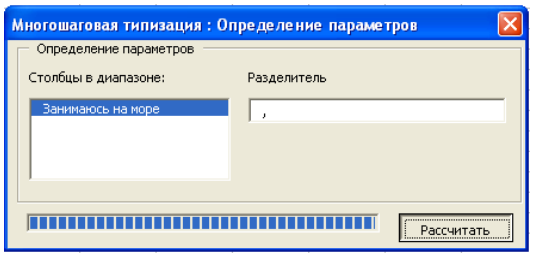

Рис. 1. Выбор признака в программе типизации

B	C	D	E
Занимаюсь на море	Класс	Подкласс	Частота
вкусно покушать			61
вязать и вышивать			4
дайвинг			58
дискотека			12
загарать и купаться			801
заниматься с детьми			8
заниматься сексом			67
заниматься спортом			72
знакомиться			8
играть в бадминтон			26
играть в баскетбол			5
играть в волейбол			367
играть в карты			23
играть в мяч			29
играть в тенис			6
играть в футбол			21
играть на гитаре			5

Рис. 2. Фрагмент таблицы типизации уникальных значений признака “занимаюсь на море” анкетного опроса по пляжно-оздоровительному отдыху

При запуске программы выводится панель управления типизацией (рис. 3). На все время активности программы типизации к таблице уникальных значений признака могут быть применимы все средства EXCEL.

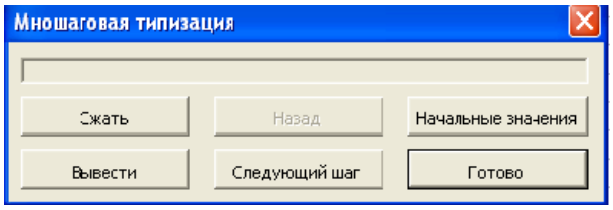

Рис. 3. Панель управления программой типизации

Первоначально этот список может содержать от 500 до 700 строк. После серии корректировок списка с целью его унификации пользователь может выполнить команду “Сжать”. По этой команде все повторяющиеся записи “сжимаются” в одну, а соответствующие частоты уникальных значений признака пересчитываются. Корректировка одной записи таблицы уникальных значений эквивалентна корректировке множества связанных с ней записей исходной таблицы данных.

При повторении нескольких циклов выполнения действий “корректировка – сжатие“, список уникальных значений быстро сокращается. По мере сокращения списка, время на обдумывание исследователем очередных корректировок возрастает, поскольку ему приходится анализировать все более и более сложные ситуации. Вместе с тем, при сокращении списка существенно сокращается время, затрачиваемое исследователем на поиск однотипных ответов.

Частоты повторения уникальных ответов (четвертый столбец таблицы) служат весьма полезной информацией для логических рассуждений исследователя. Исследователь, в первую очередь, сосредотачивает свое внимание на ответах, имеющих высокие частоты, и пытается свести к ним все остальные ответы, если это не приводит к искажению смысла ответов. В конечном итоге список удается сократить в десять и более раз, причем без искажения информации.

После завершения операции типизации признака пользователь может, либо заменить значения исходной выборки, либо, в случае сомнений в корректности действий, разместить столбец признака с замещенными значениями на новом месте. В частном случае, эта программа может быть использована для корректировки любого признака или построения частотных рядов признака. Кроме того, пользователь и сам может находить другие ситуации использования программы.

При выполнении операции типизации в полном объеме, исследователь объединяет ответы в группы, вводя названия (или номера) классов во второй столбец. В простейшем случае третий столбец просто повторяет первый.

Однако, при выполнении операции на реальных данных, возникает необходимость внесения в третий столбец значений, более общих, чем в первом столбце. В реальной ситуации могут встретиться очень близкие по смыслу, но все-таки различные ответы. Например, ответы “пробки на дорогах” и “отсутствие автостоянок”, можно было бы заменить одним обобщенным ответом – “транспортные проблемы”. Создавать два подкласса по очень близким по смыслу ответам бывает нецелесообразно, поскольку это может привести к чрезмерному количеству вариантов с крайне низкой частотой встречаемости. С другой стороны, иногда не желательно терять информацию при замене двух вариантов ответов одним обобщенным, потому что при пополнении количества данных может оказаться, что один из этих ответов достигнет такого уровня встречаемости, когда его будет целесообразно выделить, как вполне самостоятельный вариант.

Поэтому для сохранения информации “на будущее”, используется следующий подход. В строки таблицы уникальных значений, соответствующие приведенным выше ответам, вносят следующие значения: “транспортные проблемы (пробки на дорогах)” и “транспортные проблемы (отсутствие автостоянок)”, а в столбец подкласс для обоих ответов вносят

обобщенное значение “транспортные проблемы”. Определив названия классов и подклассов, исследователь может вывести результаты типизации в форме таблиц и создать новые признаки в таблице данных, составленные из значений ассоциированных с названиями классов или подклассов.

В качестве примера приведем результаты выполнения операции типизации ответов на открытый вопрос “ Чем еще любите заниматься во время отдыха на море, кроме солнечных ванн и купания?“ анкетного опроса по изучению пляжно-оздоровительного отдыха. В опросе участвовало 3361 респондент. Анкета описывается 72 признаками.

В результате выполнения операции типизации было выделено 48 различных ответов (подклассов), которые были объединены в 8 групп (табл. 1). Считается, что респонденты, дающие ответы из одной группы, обладают некоторыми общими интересами. В зависимости от того, какие ответы дал респондент, мы можем поставить ему в соответствие определенное название класса. Название класса (подкласса) служит аналогом значения, измеренного в номинальной шкале измерения.

Таблица 1

Результаты выполнения операции типизации

Класс	**Подкласс**	**Частота**
Спортсмены	спортивные игры	75
	играть в волейбол	434
	подвижные игры	144
	заниматься спортом	85
	играть в мяч	32
	играть в бадминтон	29
	играть в теннис	6
	играть в баскетбол	5
	играть в футбол	24
Увлеченные	ловить рыбу	270
	дайвинг	65
	кататься на лодке и т.п.	52
	собирать грибы и ягоды	20
	экстрим	11
	редкие увлечения	21
	активный отдых	7
Гурманы	пить пиво	69
	готовить шашлык	102
	приготовление пищи	24
	пить спиртное	92
	вкусно покушать	68
	пикник	9
	кушать сладкое и пить напитки	9

Продолжение таблицы 1

Класс	**Подкласс**	**Частота**
Лирики	читать	135
	фотографировать	14
	прогулки	97
	любоваться природой	22
	играть на гитаре	8
	слушать музыку	9
	сидеть вечером у костра	17
	строить замки из песка	12
	собирать ракушки и камни гербарий	15
	культурная программа и экскурсии	9
	разгадывать кроссворды	6
	уединение	8
Инертные	загорать купаться отдыхать	841
Общительные	общаться	59
	настольные игры	8
	заниматься сексом	71
	дискотека	14
	петь песни	16
	играть в карты	29
	знакомиться	10
	посещение кафе-баров-ночных клубов	8
Сони	спать	80
	пассивный отдых	42
Мамы	заниматься с детьми	10
	вязать и вышивать	4
Н/Д	Отсутствие данных	164

Выше мы рассмотрели использование программы типизации в простом случае. Операция типизации допускает обобщение на случай, когда респондент на один вопрос может дать не один ответ, а несколько. При этом ответы записываются в одном столбце таблицы данных, соответствующей вопросу. Несколько простых ответов разделяются каким-либо знаком (“; “ или “,”). Такой признак мы определяем как составной. Например, на вопрос о любимых занятиях в пляжной зоне респондент может ответить: “осматривать достопримечательности; играть в бадминтон; читать”. В этом случае ответ содержит три простых ответа.

Такие множественные ответы требуют разделения исследуемого признака на несколько. Для обработки таких данных применяется многошаговая типизация. Сначала производится типизация по первому ответу, затем по второму (если таковые имеются) и так далее. На каждом шаге ти-

пизации программа производит действия аналогичные рассмотренным выше для одношаговой типизации. Исправления, внесенные в данные на каждом шаге типизации, возвращаются в исходный столбец таблицы данных, либо выводятся на новом месте.

В результате типизации составного открытого ответа будет получен и составной признак в номинальной шкале измерения. Составной признак может быть получен и непосредственно при сборе первичных данных в процессе анкетирования. Составной признак можно получить, когда из списка вариантов ответа на вопрос анкеты респондент может выбрать не один вариант, а несколько. Причем, различные респонденты могут выбрать различное количество вариантов. Конечно, такой ответ можно было бы представить несколькими признаками, но такое представление далеко не всегда удобно для анализа. При построении частотного ряда простых значений, входящих в составной признак, возникает вариантность, которая не может быть разрешена с помощью стандартных средств.

Вариантность построения частотного ряда продемонстрируем на примере значений составного признака полученного в результате выполнения операции многошаговой типизации. Типизации подвергались ответы на открытый вопрос “Что омрачало ваш отдых в пляжной зоне?”

В процессе типизации было выделено 25 ответов, которые были объединены в 10 групп. Группы получили следующие названия:

1. Зеленые
2. Урбанисты
3. Нелюдимые
4. Интеллигенты
5. Автомобилисты
6. Студенты
7. Привередливые
8. Нетерпимые
9. Оптимисты
10. Равнодушные

В результате замены типовых значений названиями групп был получен новый составной признак. Фрагмент столбца значений преобразованного признака представлен в таблице 2. Если определен ограниченный список групп, то составные ответы можно представить в числовой форме. Обобщенная форма записи составного признака в числовой форме представлена в таблице 3.

Таблица 2

Значения составного признака после замены ответов названиями групп

Номер анкеты	**Значения составного признака**
1	Зеленые, Зеленые
2	Зеленые, Зеленые, Привередливые
3	Интеллигенты, Нелюдимые
4	Автомобилисты, Зеленые
5	Нелюдимые
…	
n=3361	Интеллигенты, Зеленые

Таблица 3

Числовая форма представления составного ответа

Номер анкеты	Номер группы ответов						
	1	2	3	…	j	...	$к$
1	r_{11}	r_{12}	r_{13}		r_{1j}		r_{1k}
2	r_{21}	r_{22}	r_{23}		r_{2j}		r_{2k}
3	r_{31}	r_{32}	r_{33}		r_{3j}		r_{3k}
…							
i	r_{i1}	r_{i2}	r_{i3}		r_{ij}		r_{ik}
...							
n	r_{n1}	r_{n2}	r_{n3}		r_{nj}		r_{nk}

В таблице 3 приняты следующие обозначения:

r_{ij}- количество простых ответов составного признака i- ой анкеты отнесенных к группе с номером j;

i – номер анкеты i=1,2,3,…n;

j номер группы ответов j=1,2,3,…k.

По данным таблицы 3 можно построить два варианта или модификации частотных рядов. Частоту встречаемости j – ого значения признака можно рассчитать по формуле:

$$P_j^{(1)} = \frac{\sum_{i=1}^{n} r_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}} \qquad (1)$$

и по формуле:

$$P_j^{(2)} = \frac{\sum_{i=1}^{n}\left(\frac{r_{ij}}{\sum_{j=1}^{k} r_{ij}}\right)}{n} \qquad (2)$$

Обе эти формулы дают значения, отвечающие основному свойству частотного ряда:

$$\sum_{j=1}^{k} P_j^{(1)} = \sum_{j=1}^{k} P_j^{(2)} = 1 \qquad (3)$$

В каждом конкретном случае частотные ряды, рассчитанные по формулам (1) и (2), могут существенно отличаться. То есть, для составного признака имеет место вариантность частотного ряда.

Предпочтение тому или иному варианту отдается в зависимости от того, какой содержательный смысл имеют значения составного признака. Если значения, как в нашем примере, имеет смысл типа личности, то встречаемость в одной строке исходной таблицы (табл. 2) нескольких раз-

личных значений мы можем интерпретировать как то, что конкретный респондент обладает чертами сразу нескольких типов личности. В этом случае для расчета частотного ряда предпочтительней использовать формулу (2).

Рассмотрим другой случай, приводящий к составному ответу. Например, если мы спрашиваем респондента о том, какие виды развлекательно-оздоровительных учреждений он посещает, то простые ответы “ресторан” “фитнес-клуб” целесообразно учитывать по первой схеме. То есть такой потребитель дает нагрузку двум различным типам предприятий.

С формальной точки зрения составные ответы в двух рассмотренных случаях тоже имеют различия. В первом случае r_{ij} может принимать значения 0,1,2,3, …, а во втором только значения 0 и 1.

Программные модули построения модифицированных частотных рядов по составным признакам также включены в разработанный нами специализированный пакет обработки анкетных данных. Кроме того, пакет включает программные модули, позволяющие преобразовывать составные признаки к простым и обратно.

К числу достоинств программных модулей мы относим то, что даже при очень больших выборках они позволяют получать результаты в реальном времени, что открывает большие возможности для экспериментальной работы исследователя.

Расчеты на реальных данных показали очень высокую устойчивость числовых характеристик частотных рядов, построенных по данным, полученным в результате типизации ответов на открытые вопросы. Поэтому эти данные могут выступать в роли характеристик исследуемых совокупностей. Результаты типизации могут быть с успехом использованы для анализа структуры потребителей товаров и услуг или сегментации рынка.

Наличие средств по обработке открытых вопросов обеспечивает широкому кругу исследователей новые возможности сбора первичного материала методом анкетного опроса.

Используемая литература

1. Мартышенко С.Н. Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследованиях / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Вестник ТГЭУ. 2006. № 2 С. 91–103.
2. Орлов А.И. Нечисловая статистика. М.: МЗ-Пресс, 2004. – 513 с.
3. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М.: Научный мир, 2000. – 352с.