ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В учебном пособии изложен курс лекций по анализу данных, читавшийся автором в течение ряда лет студентам-социологам II и III курса Государственного университета — Высшей школы экономики (ГУ ВШЭ) (бакалавриат направления «Социология»).

В книге рассматриваются методы, используемые социологом на практике: построение и анализ одномерных и двумерных частотных таблиц; анализ взаимосвязи качественных и количественных переменных с помощью теста Стьюдента и модели однофакторного дисперсионного анализа; построение моделей регрессии; поиск латентных переменных методами факторного анализа, главных компонент, многомерного шкалирования; получение многомерных группировок с помощью кластерного анализа. Подробно описывается, каким образом эти методы могут быть реализованы с помощью пакета SPSS — одной из самых распространенных в мире систем статистической обработки данных социальных исследований.

В последние годы появился целый ряд работ, посвященных описанию того, как можно анализировать статистические данные с помощью пакета SPSS[1]. Среди них некоторые адресованы социологам.

[1] Бююль А., Цёфель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. М.: DiaSoft, 2002; Наследов А.Д. SPSS: компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005; Плис А.И., Сливина Н.А. Практикум по прикладной статистике в среде SPSS. Ч. 1. Классические процедуры статистики. М.: Финансы и статистика, 2004; Ростовцев П.С., Ковалева Г.Д. Анализ социологических данных с применением статистического пакета SPSS: Учеб.-метод. пособие. Новосиб. гос. ун-т, 2001; Таганов Д.Н. SPSS: статистический анализ в маркетинговых исследованиях. СПб.: Питер, 2003; Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере / Под ред. В.Э. Фигурнова. М.: ИНФРА-М, 2002.

Тем не менее, предлагаемая книга имеет шанс занять достойное место в отечественной литературе, на наш взгляд, потому, что автор аккумулировал свой богатый опыт исследователя-социолога, специалиста по анализу данных и педагога, долгие годы работавшего со студентами-социологами и обладавшего талантом хорошо объяснять содержательную сущность математических построений.

В книге, во-первых, каждый из методов анализируется с точки зрения возможности решения тех или иных социологических задач. Приводятся многочисленные примеры использования рассматриваемых методов для анализа данных конкретных социологических исследований, с соответствующей точки зрения изучается специфика каждого метода. Обращается внимание на особенности интерпретации результатов анализа социологических данных. Тщательно разбираются ограничения каждого метода, по мере возможности даются рекомендации по их преодолению.

Во-вторых, при рассмотрении любого метода подробно рассматривается, каким образом на каждом шаге его реализации может использоваться пакет SPSS.

Отметим, что, читая отраженный в книге курс, автор для успешного освоения студентами технических приемов работы с компьютерными программами после каждой лекции предлагал слушателям небольшое задание для самостоятельной работы. Выполнение таких заданий контролировалось на семинарских занятиях. Практика показала эффективность подобного подхода: на базе работы с конкретными социологическими данными у студентов формировались практические навыки использования компьютерных программ при решении социологических задач.

Хочется надеяться, что социолог, прочитавший книгу, при решении стоящей перед ним задачи сумеет выбрать наиболее адекватный метод, определить и обосновать вид соответствующей математической модели, проанализировать (и преодолеть) ее ограничения, выполнить расчеты модели на компьютере, проанализировать математико-статистический смысл полученных результатов, дать соответствующую социологическую интерпретацию.

Книга рассчитана на студентов, прослушавших базовые курсы математики (математический анализ и линейная алгебра), информатики, теории вероятностей и математической статистики, а также основ социологии и методов социологических исследований. Кроме того, может быть полезна социологам для эффективного анализа имеющейся у них информации.

В подготовке книги принимали участие сотрудники и студенты кафедры методов сбора и анализа социологической информации факультета социологии ГУ ВШЭ.

1
глава

ПОДГОТОВКА К АНАЛИЗУ ДАННЫХ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

1.1

Социальное исследование и анализ данных: основные понятия

Анализ информации, собираемой в процессе эмпирических социологических исследований, представляет собой не просто совокупность технических приемов и методов, позволяющих в той или иной форме визуализировать полученные данные. Анализ данных является ключевым этапом всего исследования, в ходе которого происходит непосредственная проверка соответствия собранной информации тем моделям социальных явлений, которые, явно или латентно, имеются у социологов. И более того, в ходе анализа формулируются и проверяются новые модели, адекватно отражающие те закономерности, которые есть в собранных данных.

Очевидно, что в случае простой визуализации собранной информации мы имеем дело лишь с *обработкой* социологических данных. Если ставятся задачи построения определенной модели изучаемого социального явления и проверки соответствия этой модели имеющимся данным, можно говорить именно об *анализе* данных.

В ходе как обработки, так и анализа данных часто используют одни и те же технические и математические приемы, однако с гносеологической точки зрения это два разных подхода к данным. В первом случае социолог использует стандартный набор средств (как правило — это одномерные распределения, таблицы, гистограммы и графики) для наиболее наглядной демонстрации полученных данных, которые, при удачном подборе технических средств, вроде бы говорят сами за себя. Во втором случае исследователь выдвигает определенную модель социального явления, демонстрирует соответствие (либо противоречие) данных этой модели и ведет дальнейшую разработку именно модели, отвлекаясь от самих данных.

При работе с социологическими данными используются два основополагающих понятия:

- единица анализа (анкета, случай);
- переменная.

Единица анализа — это элементарная, единичная часть объекта исследования. В большинстве случаев единица анализа совпадает с единицей наблюдения, т.е. с тем объектом, о котором непосредственно получают информацию в ходе сбора данных. В социологии, как правило, этой единицей является отдельный респондент. Однако это не всегда так. Например, объектом изучения социолога может выступать семья как целостная единица и, следовательно, она выступает единицей анализа в исследовании. Единицами же наблюдения выступают члены семей, т.е. отдельные респонденты, о которых, собственно, и собирается информация. Преобразование информации, собранной о единицах наблюдения, в информацию о единицах анализа является самостоятельным и не только техническим этапом исследования.

Переменная — это элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа. Простейшими переменными являются, скажем, пол или зарплата респондента. Ключевыми характеристиками переменной является то, что, с одной стороны, для каждой единицы анализа она имеет одно, вполне определенное значение, а с другой стороны — то, что не все единицы анализа имеют одинаковое значение переменной.

1.2

Представление данных в пакете SPSS

Матрица, в которой представляются данные в программной системе SPSS, изображена на рис. 1.1. Редактор данных состоит из двух частей: таблицы для работы собственно с данными (рис. 1.2) и таблицы работы с переменными (рис. 1.3).

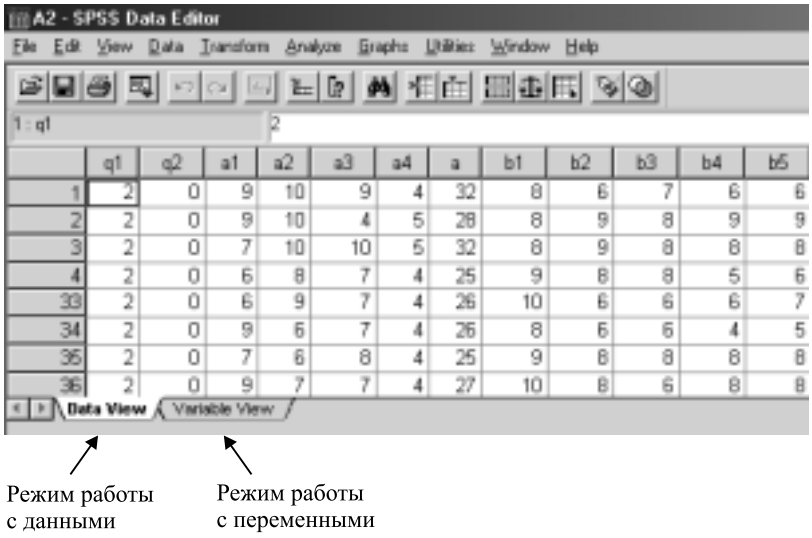


Рис. 1.1. Представление данных в пакете SPSS

Каждая строка в матрице данных содержит информацию по одной единице анализа. В примере (см. рис. 1.1) в качестве единицы анализа выступает анкета, содержащая ответы одного респондента. Все единицы анализа в матрице данных автоматически нумеруются. Номера располагаются в первой колонке матрицы данных, в остальных колонках — соответствующие значения переменных.

Прежде всего рассмотрим простейшие количественные методы анализа данных. В зависимости от решаемых задач разделим их на три основных типа.

1. Одномерный описательный анализ раскрывает некоторые характеристики частотных распределений.

2. Двумерный описательный анализ связан с описанием формы и силы взаимосвязи между переменными, а также со сравнением значений некоторой переменной в разных социальных группах.

3. Объяснительный анализ направлен на выявление силы влияния переменных друг на друга.

1.3

Построение частотных распределений

Анализ частотных распределений результатов количественного социологического исследования — это первый шаг при обработке собранной информации. Во многих случаях этот анализ не является, строго говоря, анализом данных, а выполняет функции получения общих представлений об изучаемых социальных группах.

Первый шаг одномерного описательного анализа для объяснения какого-то явления — его описание. Результаты любого массового опроса содержат ответы большого числа респондентов на широкий круг анкетных вопросов. Даже в рамках только одного вопроса анкеты объем исходной информации достаточно велик для того, чтобы можно было охватить его одним взглядом и каким-то образом суммировать. Именно задачу сжатия исходной информации, компактного ее представления для дальнейшего осмысления и решают методы одномерного описательного анализа.

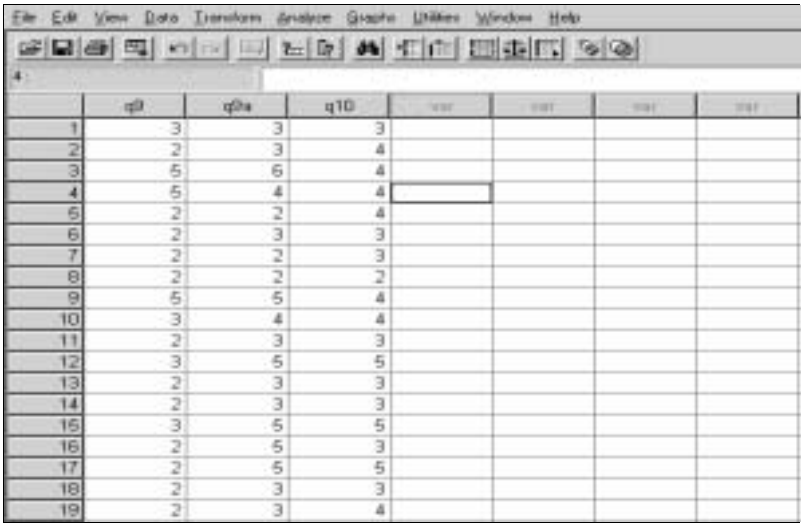
Одномерный описательный анализ решает поставленную задачу взаимодополняющими методами:

- построения частотных распределений;
- графического представления поведения анализируемой переменной;

• получения статистических характеристик распределения анализируемой переменной.

В табл. 1.1 представлен фрагмент данных по результатам социологического опроса[1].

Таблица 1.1. Фрагмент матрицы данных из трех переменных в формате SPSS, содержащий результаты социологического опроса



	q9	q9a	q10	var	var	var	var
1	3	3	3				
2	2	3	4				
3	5	6	4				
4	5	4	4				
5	2	2	4				
6	2	3	3				
7	2	2	3				
8	2	2	2				
9	5	5	4				
10	3	4	4				
11	2	3	3				
12	3	5	5				
13	2	3	3				
14	2	3	3				
15	3	5	5				
16	2	5	3				
17	2	5	5				
18	2	3	3				
19	2	3	4				

Переменная q9, представленная во второй колонке матрицы, содержит ответы респондентов на вопрос анкеты:

q9 Что вы могли бы сказать о своем настроении в последние дни?

1. Прекрасное настроение.
2. Нормальное, ровное состояние.
3. Испытываю напряжение, раздражение.
4. Испытываю страх, тоску.
5. Затрудняюсь ответить.

[1] Опрос проводился ВЦИОМ «Мониторинг общественного мнения» // Мониторинг общественного мнения: экономические и социальные перемены. 2002. № 6.

В матрице данных ответы представлены в виде числовых кодов. Поскольку полностью вся матрица содержит ответы 2407 респондентов, просто просмотр ответов всех опрошенных либо на экране компьютера, либо в распечатанном виде на листах бумаги не дает возможности понять, каково было настроение опрошенных. Получить обобщенную, агрегированную картину ответов на данный вопрос позволяет таблица одномерного частотного распределения, представленная в табл. 1.2.

Таблица 1.2. Одномерное частотное распределение переменной q9

	Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Прекрасное настроение	158	6,6	6,6	6,6
Нормальное, ровное состояние	1185	49,2	49,2	55,8
Испытываю напряжение, раздражение	752	31,2	31,2	87,0
Испытываю страх, тоску	163	6,8	6,8	93,8
Затрудняюсь ответить	149	6,2	6,2	100,0
Total	2407	100,0	100,0	

Построение одномерного частотного распределения в рамках пакета SPSS выполняется с помощью команды *Frequencies*, расположенной в блоке команд *Descriptives* (рис. 1.2). На рис. 1.3 представлено меню команды *Frequencies*.

Таблица 1.2 демонстрирует одномерное частотное распределение переменной q9 в том виде, как это распределение вычисляется командой *Frequencies* пакета SPSS. Рассмотрим информацию, которую дает таблица одномерного частотного распределения.

Колонка *Frequency* (частота) содержит частоты, т.е. то количество респондентов, которые выбрали тот или иной вариант ответа. Таким образом, из табл. 1.2 видно, что вариант «1» выбрали 158 респондентов, вариант «2» — 1185 респондентов и т.д. Последняя стро-

ка в табл. 1.2 — *Total* — в колонке *Frequency* показывает общее количество опрошенных, иными словами — объем выборки.

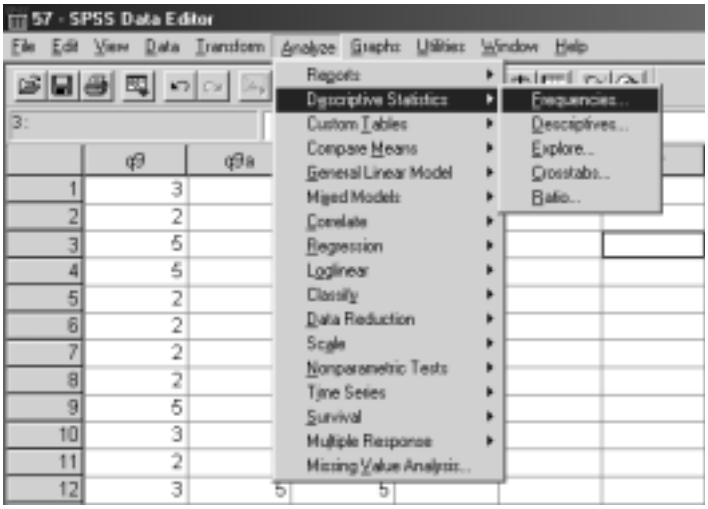


Рис. 1.2. Метод вызова команды построения одномерных частотных распределений в пакете программ SPSS

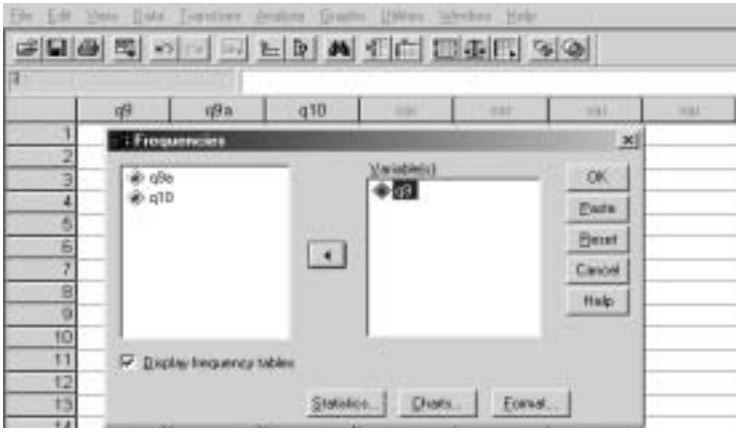


Рис. 1.3. Меню команды построения одномерных частотных распределений

Делать выводы о том, много или мало респондентов отметили при опросе ту или иную градацию в вопросе, опираясь на значения в колонке *Frequency*, невозможно, поскольку необходимо постоянно соотносить эти числа с общим количеством опрошенных. Поэтому удобнее использовать колонку *Percent* (процент), которая содержит процентные значения для каждой из частот. В результате, базируясь на значениях этой колонки, можно сказать, что более распространенным ответом является «нормальное, ровное состояние», поскольку этот вариант отметили 49,2% респондентов.

Колонка *Valid Percent* связана с такой важной в социологической практике характеристикой, как «Отсутствие ответа». Мы знаем, что в ходе любого массового опроса какая-то часть опрашиваемых не отвечает на поставленные вопросы. Причины такого рода «неответов» различны. Это и просто нежелание людей давать информацию по тем или иным показателям. Это и отсутствие собственного мнения по определенным вопросам. Возможности преодоления проблемы «неответов» на этапе сбора социологической информации достаточно подробно рассматриваются у разных авторов, однако очевидно, что эту проблему нельзя решить полностью.

На этапе работы с собранными данными проблема «неответов» может быть сформулирована следующим образом: как анализировать ту информацию, которая может быть квалифицирована как «отсутствие ответа».

Необходимо отметить, что на этот вопрос нет однозначного ответа. В зависимости от характера решаемых задач существуют разные подходы к анализу информации, которая соответствует «неответам». Отметим, что числовые коды, связанные с «неответами», называют *пропущенные данные* (Missing values).

Первый подход к рассмотрению кодов пропущенных данных рассматривает эти коды как равноправные остальным числовым кодам, которые приписаны всем другим типам ответов. Одномерное частотное распределение (см. табл. 1.2) представляет именно такой подход. Действительно, числовой код «5» приписанный варианту «Затрудняюсь ответить», т.е. фактически коду пропущенных данных, представ-

лен точно так же, как и остальные числовые коды. В результате табл. 1.2 демонстрирует нам, что затруднились ответить на поставленный вопрос 149 человек, или 6,2% общего числа опрошенных. При этом и все остальные проценты в табл. 1.2 рассчитаны от *числа опрошенных.*

Альтернативным вариантом построения таблицы одномерного частотного распределения выступает возможность исключения из дальнейшего анализа тех респондентов, которые затруднились дать ответ. Действительно, какие у нас основания рассматривать тех 149 респондентов, которые не дали ответа на поставленный вопрос точно так же, как и тех, кто дал содержательный ответ? Простейшим выходом в рамках данной модели рассуждений является приписывание коду «5» статуса пропущенных данных и исключение из дальнейшего анализа тех, кто дал такой ответ. В табл. 1.3 представлено одномерное частотное распределение, в котором коду «5» приписан статус пропущенных данных.

Таблица 1.3. Одномерное частотное распределение переменной q9

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	1. Прекрасное настроение	158	6,6	7,0	7,0
	2. Нормальное, ровное состояние	1185	49,2	52,5	59,5
	3. Испытываю напряжение, раздражение	752	31,2	33,3	92,8
	4. Испытываю страх, тоску	163	6,8	7,2	100,0
	Total	2258	93,8	100,0	
Missing	5. Затрудняюсь ответить	149	6,2		
	Total	2407	100,0		

Таблица 1.3 отличается от табл. 1.2 тем, что код «5» помечен как код пропущенных данных. Колонка *Percent*, как и раньше, содержит процентное *распределение всех опрошенных*, а колонка *Valid Percent* — процентное распределение от того числа респондентов, которые дали ответы, не помеченные кодами пропущенных данных. Иными словами, колонка *Valid Percent* представляет одномерное частотное *распределение от числа ответивших респондентов*.

Вопрос о том, какой из показателей — процент опрошенных, либо процент ответивших необходимо использовать для выявления определенных социологических закономерностей, некорректен. Оба показателя несут определенную информацию и, как правило, используются одновременно, однако их интерпретация существенно различна. Например, если в ходе опроса, за кого собираются голосовать респонденты на предстоящих выборах, мы получим, что за кандидата А собирается голосовать 20% опрошенных и 40% ответивших, то оба этих числа представляют интерес. Действительно, первое число говорит нам, что 20% общего количества взрослого населения собирается поддержать кандидата А на будущих выборах. Поскольку коды пропущенных данных в такого рода опросах получают, как правило, те респонденты, которые говорят, что не будут участвовать в выборах, то число 40% говорит нам о том, сколько процентов может набрать кандидат А в ходе голосования.

Присвоение кода пропущенных данных для переменных в пакете SPSS выполняют по таблице «Описание переменных». На рис. 1.4 приведены таблица «Описание переменных» и показатель, с помощью которого задаются коды пропущенных данных. В нашем примере у переменной q9 задан код пропущенных данных «5».

На рис. 1.5 представлено меню, которое позволяет задавать коды пропущенных данных для выбранной переменной, а также показывает три варианта задания кодов пропущенных данных:

- не определять коды пропущенных данных для переменной (*No missing values*);
- задать несколько (от 1 до 3) значений кодов пропущенных данных (по одному значению в каждом из открытых окон — *Discrete missing values*);

• задать интервал значений кодов пропущенных данных и одно значение кода пропущенных данных (*Range plus one optional discrete missing value*).

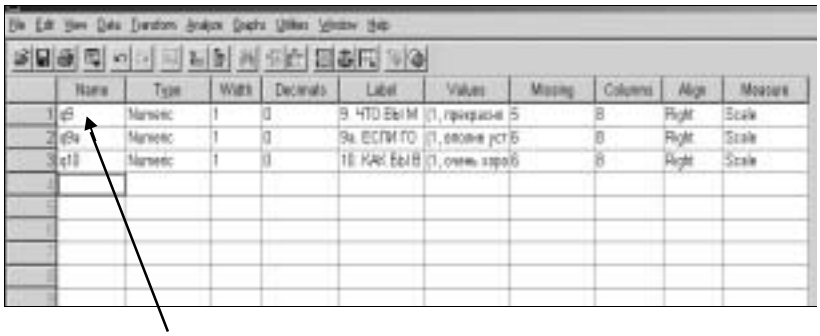


Рис. 1.4. Таблица «Описание переменных» в пакете программ SPSS

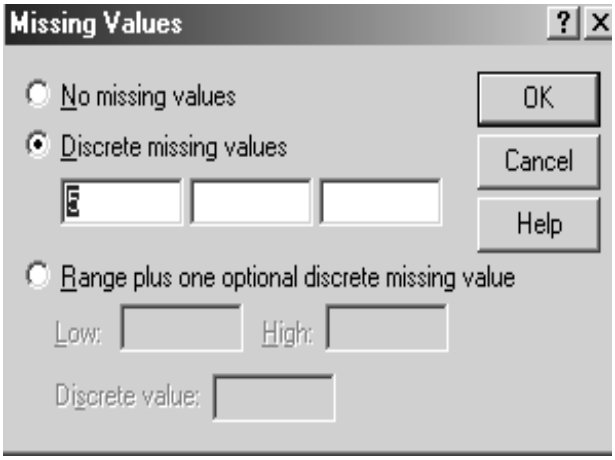


Рис. 1.5. Меню задания кодов пропущенных данных

По какой причине может потребоваться задание не одного, а нескольких кодов пропущенных данных? Эта возможность отражает реаль-

ные ситуации, по которым в ходе опроса мы нередко имеем несколько причин того, что респондент не отвечает на вопрос анкеты. Рассмотрим вопрос, в котором присутствует несколько вариантов, каждый из которых объясняет причину, почему респондент не дает ответа.

За кого из кандидатов вы голосовали на прошедших выборах?

1. За кандидата А.
2. За кандидата В.

...

7. Я не участвовал в голосовании.
8. Я голосовал, но не помню за кого.
9. Я участвовал в голосовании, но не хочу говорить, за кого отдал свой голос.

С точки зрения социологического анализа результатов голосования, коды «7», «8» и «9» должны быть определены как коды пропущенных данных, поскольку эти ответы не содержат информации о том, за кого голосовал респондент. Однако с точки зрения социологических задач весьма интересным может быть изучение, например, характеристик тех респондентов, кто не участвовал в голосовании. Для осуществления анализа этой социальной совокупности мы можем отменить задание кода «7» как кода пропущенных данных и сосредоточиться на анализе данной группы респондентов.

Третий вариант задания кодов пропущенных данных чаще всего встречается в ситуации, когда анализируемая переменная выражена количественно. Иногда в ответах респондентов на вопросы, например, о размере получаемых доходов встречаются данные, которые, строго говоря, не могут быть признаны ошибочными, однако, скорее всего, являются недостоверными. Например, если респондент сказал, что у него 15 детей или что его зарплата 5 млн. руб., эти ответы едва ли корректны. Иными словами, для многих показателей мы можем указать границы, допустимых значений, а те данные, которые выходят за эти границы целесообразно признать пропущенными данными. В меню определения кодов пропущенных данных в разделе *Range plus one optional discrete missing value* (интервал и, возможно, одно значение пропущенных данных) можно задать верхнюю и нижнюю границы

интервала, все значения внутри будут являться пропущенными данными. В этом разделе наряду с интервалом можно задать одно точное значение кода пропущенных данных.

1.4

Графическое представление поведения анализируемой переменной

Наряду с табличным представлением одномерное частотное распределение можно визуализировать в графической форме. Наиболее популярные формы — это столбиковые и круговые диаграммы. На рис. 1.6 и 1.7 представлены эти виды диаграмм для одномерного частотного распределения табл. 1.2.

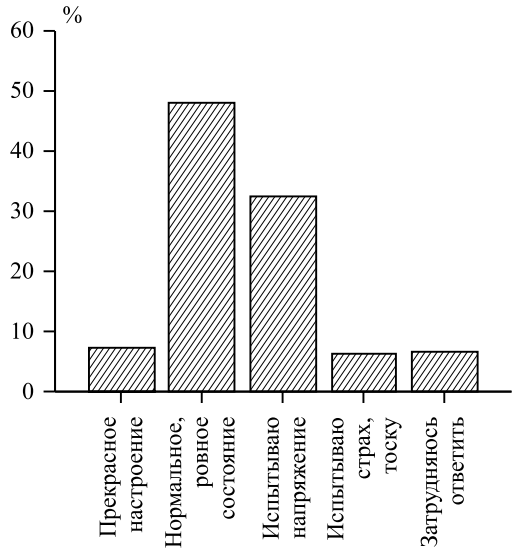


Рис. 1.6. «Что вы могли бы сказать о своем настроении в последние дни?»

Рис. 1.7. «Что вы могли бы сказать о своем настроении в последние дни?»

Диаграммы на рис. 1.6 и 1.7 построены с помощью графических возможностей пакета программ SPSS. Команды для построения графических диаграмм могут выполняться либо непосредственно из модуля вычисления одномерных частотных распределений (команда *Frequencies*), либо из специального блока команд *Graphs*, в котором представлены возможности графического анализа пакета программ SPSS.

В нижней части меню команды *Frequencies* (см. рис. 1.3) есть педаль Charts..., нажатие на которую приводит к вызову меню построения диаграмм одномерного частотного распределения (рис. 1.8).

Графические диаграммы в качестве метода построения одномерных частотных распределений повышают наглядность полученных закономерностей и могут использоваться, прежде всего, для презентации результатов социологических исследований. Какой из видов диаграмм выбрать для каждого конкретного случая — зависит от эстетических пристрастий и существенного значения не имеет.

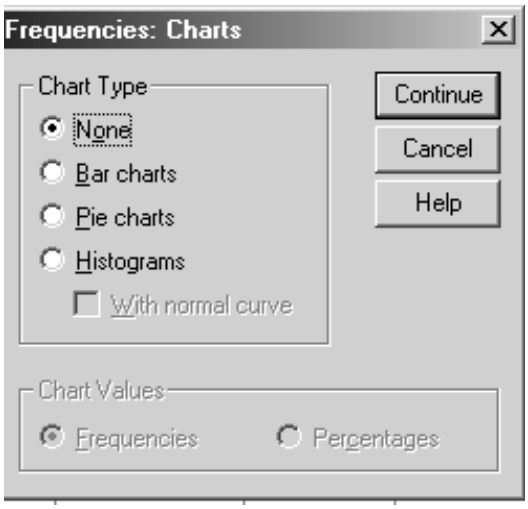


Рис. 1.8. Меню Charts команды Frequencies

1.5

Использование статистических характеристик для анализа одномерных распределений

Одной из важнейших характеристик при описании поведения отдельных переменных является показатель средней тенденции. В курсе «Методы социологического исследования» подробно обсуждаются вопросы уровней измерения, используемые в социологических анкетах, а также рассматриваются возможности применения различных мер центральной тенденции для показателей с разным уровнем измерения[2].

[2] См.: Ядов В.А. Социологическое исследование: методология, программа, методы. Самара: Изд-во Самарского ун-та, 1995. С. 98—109.

Возможности использования различных мер средней тенденции для шкал различного типа приведены в табл. 1.4.

Таблица 1.4. Возможности использования различных мер средней тенденции для шкал различного типа

№ п/п	Уровень измерения	Допустимые меры средней тенденции
1	Номинальный	Мода
2	Порядковый	Мода, медиана
3	Метрический	Мода, медиана, среднее арифметическое

Рассмотрим специфику использования мер средней тенденции для анализа социологических данных на примере среднего арифметического. Среднее арифметическое широко используется в повседневной жизни и не нуждается в дополнительных рекомендациях. Вместе с тем использование *только* среднего арифметического для описания значений переменной таит определенную опасность.

Говоря о среднем значении некоторой переменной мы, по сути дела, заменяем рассмотрение всей совокупности значений этой переменной единственным показателем, фактически предполагая, что значение этого показателя достаточно хорошо описывает поведение анализируемой переменной. Очевидно, что в данном случае среднее значение выступает в качестве определенной модели значений переменной.

Несомненно, что среднее арифметическое переменной представляет совокупность значений этой переменной неполно и с возможными ошибками. Зная, например, среднее значение зарплаты среди совокупности опрошенных, мы не можем достаточно точно определить зарплату того или иного респондента. Только в том случае, когда все значения переменной одинаковы, среднее значение абсолютно точно отражает поведение переменной. Во всех других случаях среднее арифметическое как модель переменной является моделью неточной. Следовательно, для нас важно знать не только значение данной модели, но и степень точности, качества этой модели.

Рассмотрим данные о заработной плате пяти респондентов, полученные в ходе социологического исследования (табл. 1.5).

Таблица 1.5. Данные о средней заработной плате, среднее значение заработной платы, расхождение среднего и фактических данных

№ п/п	Значение заработной платы, руб.	Среднее значение, руб.	Расхождение реальной зарплаты и среднего значения, руб.
1	17 000	15 500	1500
2	13 000	15 500	–2500
3	18 000	15 500	2500
4	15 000	15 500	500
5	14 500	15 500	–1000

Данные, приведенные в табл. 1.5, можно представить в виде условной формулы:

Реальные данные = Модель + Остаток.

Расхождение реальных данных и модели в этой формуле называется остатком.

В каком случае модель средней зарплаты будет с небольшой погрешностью описывать реальные данные? Ключевым вопросом при анализе данных с помощью какой бы то ни было модели является оценка того, насколько хороша модель. Остатки дают нам эффективный инструмент для оценки качества модели: очевидно, что модель тем лучше, чем меньше остатки.

Таким образом, наряду со средней характеристикой, которая удобна тем, что дает нам картину (вернее, часть картины) поведения значений переменной, целесообразно иметь и еще одно число, которое оценивало бы качество средней как модели. Функции такой характеристики выполняют *меры разброса*, наиболее известна среди них *дисперсия*.

Фактически дисперсия представляет собой не что иное, как сумму квадратов остатков, деленную на количество наблюдений:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (1.1)$$

где x_i — значение переменной x для i-го респондента; $\bar{x}$ — среднее значение переменной x; n — количество опрошенных респондентов.

Недостатком дисперсии является то, что эту величину трудно оценить интуитивно. Данные, представленные в табл. 1.5, имеют понятные нам единицы измерения — рубли. Поэтому мы сразу можем оценить, что за величина остатка, скажем, у респондента 4 — 500 руб. Понятна нам и размерность среднего показателя — 15 500 руб. Мы можем интерпретировать это значение, соотнося его с нашим знанием социальной действительности.

В то же время значение дисперсии для данных табл. 1.5 составляет 4 000 000. Едва ли мы можем, хотя бы на качественном уровне, оценить, большая эта величина или маленькая. Это значение не дает нам ответа на главный вопрос — хороша ли наша модель среднего арифметического, т.е. средней зарплаты. Причина того, что дисперсия плохо приспособлена для ответа на вопрос о качестве модели среднего, в том, что остатки берутся в квадрате. Для того чтобы преодолеть это затруднение, используют два производных от дисперсии показателя — стандартное отклонение и стандартная ошибка среднего.

Стандартное отклонение — это корень квадратный из дисперсии. Стандартное отклонение для данных табл. 1.5 — 2000.

$$S = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{n}}. \qquad (1.2)$$

Стандартная ошибка среднего (с.о. $\bar{x}$) тоже широко используется для решения задачи оценки качества среднего как модели с несколько иной стороны: она дает возможность соотнести величину $\bar{x}$ с генеральным математическим ожиданием. Последнее с вероятностью 0,95 лежит в интервале $(\bar{x} \pm 2\text{с.о.}\bar{x})$.

$$\text{с.о.}\overline{x} = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}{n}. \qquad (1.3)$$

По табл. 1.5 значение стандартной ошибки среднего составляет 894. Таким образом, можно утверждать, что с вероятностью 0,95 математическое ожидание зарплаты должно лежать в интервале 15 500 ± ± 2 × 894, или от 13 712 до 17 288 руб. (см. п. 1.7).

Подводя итог, необходимо подчеркнуть, что использование среднего арифметического без указания одного из показателей качества среднего как модели (дисперсии, стандартного отклонения, либо стандартной ошибки среднего) не дает возможности удовлетворительной интерпретации полученного среднего.

Проведенные рассуждения о необходимости дополнения характеристики средней тенденции показателем качества этой модели справедливо и в отношении тех переменных, которые измерены на номинальном или порядковом уровне. Для номинальных переменных мерой центральной тенденции может выступать только мода, т.е. наиболее часто встречающееся значение переменной. Мода не имеет какого-то показателя разброса. Определенной характеристикой может считаться лишь само процентное значение модальной величины. В качестве примера рассмотрим табл. 1.6, в которой приведено одномерное частотное распределение респондентов, проживающих в населенных пунктах разного типа.

В табл. 1.6 модальным значением является «2». Тот факт, что на эту градацию приходится 53,7% всех опрошенных респондентов, говорит о том, что на все остальные градации приходится лишь 46,7%, что может указывать на разброс значений. Однако данное указание достаточно слабо, поскольку не показывает, как именно разбросаны данные по другим градациям анализируемой переменной.

Для переменных, измеренных на порядковом уровне, основной мерой центральной тенденции является медиана. Рассчитаем медиану для переменной q23: *Насколько вы удовлетворены состоянием своего здоровья?*, которая фиксирует ответы респондентов по 7-балльной порядковой шкале (табл. 1.7).

Таблица 1.6. Одномерное частотное распределение переменной CITY «Тип населенного пункта»

№ п/п	Населенный пункт	Frequency	Percent	Valid Percent	Cumulative Percent
1	Москва	520	21,5	21,5	21,5
2	Областной центр	1300	53,7	53,7	75,2
3	Малый город в области	350	14,5	14,5	89,7
4	Сельский населенный пункт	250	10,3	10,3	100,0
Total		2420	100,0	100,0	

Медиана является такой точкой на шкале, которая делит всю совокупность опрошенных на две равных части — тех, кто отметил градации меньше этой точки (либо равные ей), и тех, кто отметил градации больше этой точки. Из табл. 1.7 видно, что в вопросе q23 градации 1, 2, 3 и 4 отметили 50,4% респондентов, и, следовательно, градация «4» является медианой.

Таблица 1.7. Одномерное частотное распределение переменной q23

№ п/п		Frequency	Percent	Valid Percent	Cumulative Percent
1	Полностью удовлетворен	336	12,2	12,2	12,2
2		355	12,9	12,9	25,1
3		388	14,1	14,1	39,2
4		308	11,2	11,2	50,4
5		322	11,7	11,7	62,1
6		360	13,1	13,1	75,2
7	Совершенно неудовлетворен	685	24,9	24,9	100,0
Total		2754	100,0	100,0	

Наиболее распространенным показателем, характеризующим разброс значений переменной, измеренной на порядковом уровне, является *квартильное отклонение*. Чтобы понять смысл этого показателя, необходимо уяснить значение понятия *квартиля.*

Квартиль является естественным развитием медианы, с той разницей, что квартильное разбиение делит всех респондентов не на 2, а на 4 части. Первый квартиль — это такая точка на шкале, значения меньше (либо равные) которой отметили 25% опрошенных. Второй квартиль — точка, меньше которой отметили 50% опрошенных (следовательно, второй квартиль совпадает с медианой). Наконец, третий квартиль — точка, градации меньше которой отметили 75% опрошенных.

В примере табл. 1.7 первый квартиль — это градация «2» переменной q29, поскольку градации «1» или «2» отметили 25,1% опрошенных. Второй квартиль (медиана) — «4», а третий квартиль — градация «7». Квартильное отклонение — это разница между третьим и первым квартилями. В рассматриваемом примере квартильное отклонение равно 5. При том что в целом рассматриваемая переменная q23 имеет 7 градаций, квартильное отклонение, равное 5, может рассматриваться как достаточно большое, если рассматривать шкалу как метрическую, можно сделать вывод, что модель средней тенденции (в данном случае — медиана) неточно отражает поведение переменной, поскольку много респондентов имеют значения переменной, существенно отличающиеся от медианы.

Обдумывая логику разбиения совокупности значений переменной на 2 (медиана), либо на 4 (квартили) равнонаполненных части, вполне можно поставить задачу разбиения и на 5, и на 10, и вообще на любое количество равных частей. Действительно, при анализе социологических данных иногда используются *квинтильное* (на 5 равных частей) и *децильное* (на 10 равных частей) разбиения. Соответственно применительно к таким разбиениям можно использовать такие меры разброса, как квинтильное и децильное отклонения.

Вызов блока вычисления мер средней тенденции и разброса в рамках команды построения одномерных частотных распределений проводится с помощью кнопки *Statistics* в нижней части главного меню команды *Frequencies* (см. рис. 1.3). Нажатием этой кнопки вызывают на экран меню Statistics (рис. 1.9).

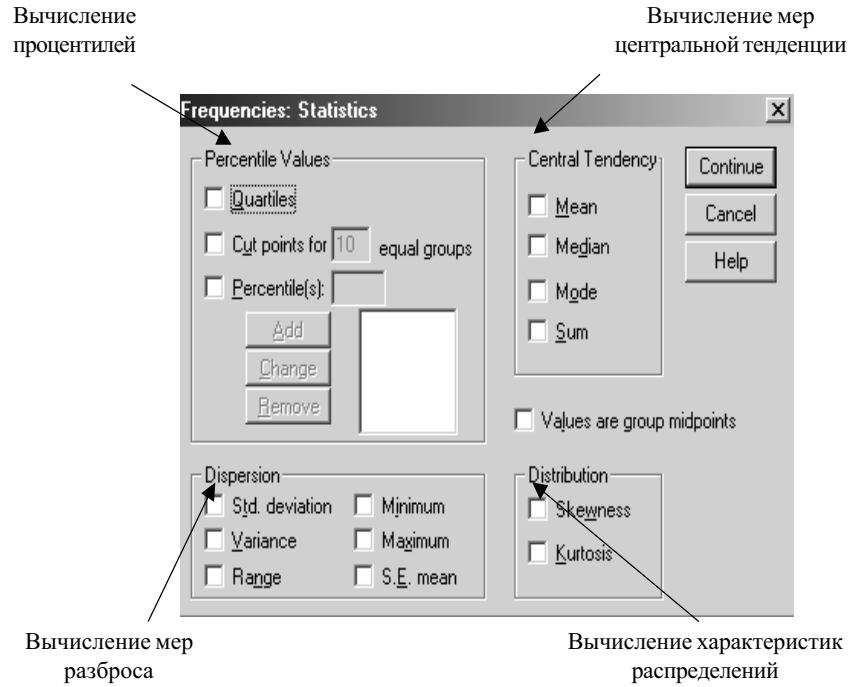


Рис. 1.9. Меню Statistics команды вычисления одномерных частотных распределений

Меню Statistics состоит из четырех отдельных блоков:

- вычисление мер центральной тенденции (Central Tendency);
- вычисление процентилей (квартили, квинтили и т.п.) (Percentile Values);
- вычисление мер разброса (Dispersion);
- вычисление характеристик распределений (Distribution).

Выбор необходимых окон в каждом из блоков приводит к вычислению соответствующих статистических показателей. Отметим, что в рамках меню Statistics команды *Frequencies* невозможны вычисления показателей квартильного (квинтильного, децильного и т.п.) отклонения. Вычисляются только сами точки процентильного разбиения.

Проиллюстрируем это вычислением квартилей для переменной q23, одномерное частотное распределение см. в табл. 1.7. В случае выбора в меню Statisics окна *Quartiles* вычисляются следующие статистики (табл. 1.8). Данные табл. 1.8 представляют все необходимое для вычисления квартильного отклонения.

Таблица 1.8. Квартильное разбиение для переменной «Насколько вы удовлетворены состоянием своего здоровья?»

N	Valid	2754
	Missing	0
Percentiles	25 50 75	2,00 4,00 7,00

Отметим, что при вычислении всех статистических характеристик только значения, не отмеченные кодами пропущенных данных.

Полезным и нередко используемым показателем при анализе количественных переменных является *децильное отношение*. Продемонстрируем использование данного показателя на примере. В ходе социологического исследования, проведенного в сентябре 2003 г. ВЦИОМ, респондентам, в частности, задавался вопрос о размере их заработной платы на основном месте работы. При анализе данного показателя возникла потребность изучить, насколько высока неоднородность значений получаемой респондентами заработной платы.

В качестве первого шага для решения этой задачи было построено децильное разбиение исследуемого показателя (табл. 1.9).

О чем говорят материалы табл. 1.9? О том, что заработную плату до 1800 руб. получают 10% опрошенных (граница первого дециля), а также о том, что 10% опрошенных получают зарплату в размере 15 000 руб. и выше (граница десятого дециля).

Децильное отношение — это отношение десятого дециля к первому. Этот показатель демонстрирует, во сколько раз больше получают 10% наиболее высокооплачиваемых респондентов по сравне-

нию с 10% наименее оплачиваемых. В нашем примере децильное отношение составляет 8,3, что показывает степень неоднородности заработной платы.

Таблица 1.9. Децильное разбиение для переменной «Размер вашего заработка за последний месяц»

N	Valid	1079
	Missing	0
Percentiles	10	1800
	20	3000
	30	3600
	40	4500
	50	6000
	60	7500
	70	9000
	80	11 100
	90	15 000

1.6

Стандартизация показателей

Одной из задач, возникающих при одномерном анализе социологических данных, является сопоставление значения определенной переменной для конкретного респондента со средним значением этой переменной в какой-то социальной группе. Например, если результаты опроса показали, что некий респондент за последний месяц потратил 70 руб. на покупку хлеба, и не зная средней величины затрат на покупку данного вида товаров в том регионе, где проживает респондент, мы не можем сказать, много или мало денег потратил респондент на хлеб. Величина «70 рублей» может быть осознана и проинтерпретирована только в сравнении с затратами других респонден-

тов. Для того чтобы сразу оценить относительную величину того или иного количественного показателя для конкретного респондента, используется метод стандартизации исходных данных.

Существует несколько различных подходов к стандартизации данных, но самый распространенный — это так называемая *Z*-стандартизация. Вычисление стандартизованной величины Z_{xi} для значения переменной x для i-го респондента проводится по формуле

$$Z_{xi} = \frac{x_i - \bar{x}}{S}, \qquad (1.4)$$

где x_i — значение переменной для i-го респондента; $\bar{x}$ — среднее значение переменной x; S — стандартное отклонение для переменной x.

Значение показателя Z_{xi} для i-го респондента более информативно с точки зрения задачи относительного положения данного респондента, чем значение исходной переменной x_i. Действительно, из формулы (1.4) следует, что если для i-го респондента Z_{xi} положительно, данный респондент имеет значение переменной x_i большее, чем средний опрошенный респондент. Таким образом, знак Z_{xi} сразу говорит нам о положении респондента (по переменной x) относительно других опрошенных.

После того как мы выяснили, большее или меньшее значение по переменной x имеет данный респондент по сравнению с другими опрошенными, необходимо узнать, насколько это значение больше или меньше, чем у других респондентов. Из свойств стандартного нормального распределения следует, что 68% Z_{xi} должны лежать в интервале от –1 до 1, а 95% — в интервале от –2 до 2. Таким образом, если по модулю значение Z_{xi} меньше единицы, мы можем сказать, что значение переменной x для данного респондента вполне типично. Если значение Z_{xi} по модулю находится от 1 до 2, можно говорить, что данный респондент по рассматриваемому показателю значительно отличается от среднего респондента. Наконец, если Z_{xi} по модулю превосходит 2, можно утверждать, что данный респондент резко отличается от среднего[3].

[3] Все последние утверждения, строго говоря, справедливы лишь в ситуации, когда распределение исходной переменной x не сильно отличается от нормального. Вместе с тем практика показывает, что в абсолютном большинстве случаев это именно так.

В блоке команд *Descriptives statistics* есть команда, с помощью которой можно провести *Z*-стандартизацию для отобранных переменных. Это команда *Descriptives* (рис. 1.10).

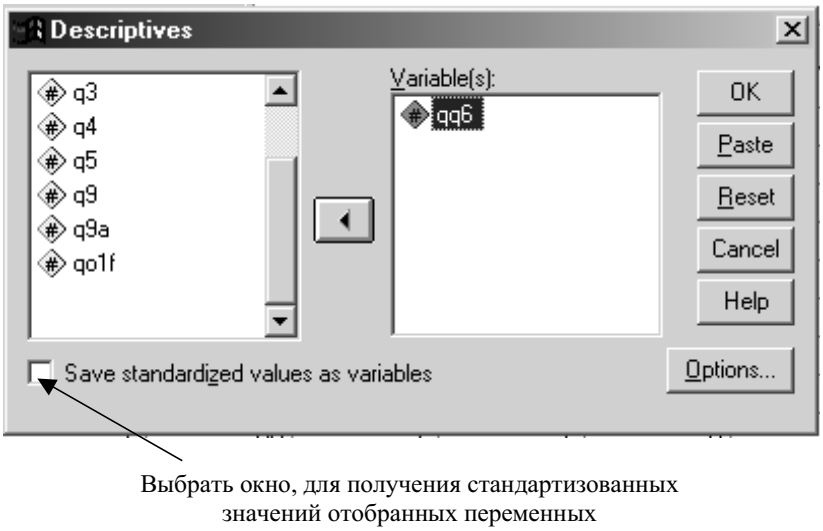
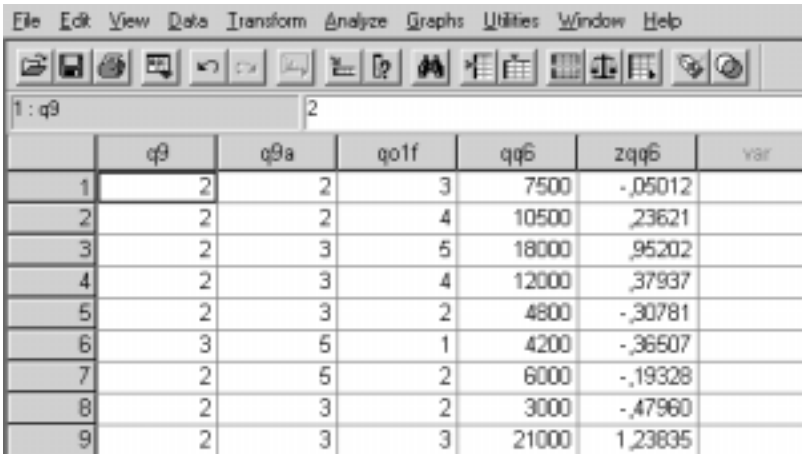


Рис. 1.10. Главное меню команды Descriptives

Команда *Descriptives* в значительной степени дублирует функции команды *Frequencies*, поскольку отвечает за вычисление одномерных статистических характеристик (мер средней тенденции и мер разброса) для выбранных переменных. Важной задачей, которую решает команда *Descriptives*, является вычисление *Z*-стандартизованных значений. В нижней части главного меню команды находится окно, при выборе которого SPSS автоматически вычисляет стандартизованные значения для выбранных переменных.

На рис. 1.11 приведена матрица данных, которая получается после выполнения стандартизации переменной qq6[4].

[4] Переменная qq6 содержит ответы на вопрос: каков был размер вашего заработка, доходов от основной работы, полученных в прошлом месяце (после вычета налогов).



	q9	q9a	qo1f	qq6	zqq6	var
1	2	2	3	7500	-,05012	
2	2	2	4	10500	,23621	
3	2	3	5	18000	,95202	
4	2	3	4	12000	,37937	
5	2	3	2	4800	-,30781	
6	3	5	1	4200	-,36507	
7	2	5	2	6000	-,19328	
8	2	3	2	3000	-,47960	
9	2	3	3	21000	1,23835	

Рис. 1.11. Матрица данных, содержащая значения переменной qq6 и стандартизованной переменной zqq6

Переменная zqq6 представляет собой стандартизованное значение переменной qq6. Использование стандартизованной переменной позволяет сказать, что величина зарплаты у первого респондента приблизительно равна среднему значению по массиву опрошенных. А вот размер заработной платы у респондента номер 9 значительно выше, чем у среднего респондента.

Использование стандартизованных переменных весьма полезно и при решении задачи сопоставления показателей, измеренных в разных единицах. Например, в нашем распоряжении есть данные по опросам в России и США, и получается, что у российского респондента А средняя зарплата составляет 9000 руб. в мес., а у американского респондента В — 2000 долл. в мес. Очевидно, что, не зная значений средней зарплаты в России и США, мы не можем сказать, выше ли респондент А респондента В, с точки зрения средней заработной платы, в их социальном кругу.

Если у нас есть возможность сопоставлять не исходные данные о величинах зарплат, а соответствующие стандартизованные показатели, мы легко можем ответить на поставленный вопрос.

1.7

Интервальное оценивание

Одномерное частотное распределение позволяет констатировать определенные закономерности в той совокупности респондентов, которые были опрошены в ходе проведенного исследования. Однако объектом социологического исследования выступает, в абсолютном большинстве случаев, не та совокупность респондентов, которая непосредственно опрашивается, а какая-то социальная либо социально-демографическая группа. Опрошенные респонденты выступают лишь как представители этой группы, как выборка, которая призвана репрезентировать поведение группы в целом. Поэтому возникает закономерный вопрос: как соотносится одномерное распределение, характеризующее поведение той или иной переменной в выборочной совокупности, с поведением этой переменной во всей анализируемой социальной общности? Иными словами, как можно перенести результат, полученный для выборки, на всю изучаемую генеральную совокупность?

Поскольку размер обследованной выборочной совокупности существенно меньше, чем генеральная совокупность, то перенесение результатов с выборочной совокупности на генеральную возможно лишь с определенной точностью. Иными словами, если в ходе опроса получено, что в выборочной совокупности 6,9% опрошенных ответили, что они «в целом довольны своей жизнью», это вовсе не значит, что во всей генеральной совокупности своей жизнью довольны именно 6,9% населения. Выборочный метод дает нам правило, которое позволяет, зная значение определенного параметра в выборочной совокупности, оценить возможное значение этого параметра в генеральной совокупности[5].

[5] Подробно вопросы генерализации результатов социологических опросов см.: Батыгин Г.С. Лекции по методологии социологических исследований. М.: Аспект-Пресс, 1995. С. 145—189.

Теоремы математической статистики говорят нам, что если выборка исследования реализуется с соблюдением определенных требований, результаты, полученные на выборке, могут быть перенесены на генеральную совокупность доверительных интервалов. Таким образом, если в выборочной совокупности оказалось 6,9% респондентов, довольных своей жизнью, в генеральной совокупности таких респондентов будет (6,9 ± Δ)%. Величина Δ называется максимальной ошибкой выборки, а интервал (6,9 – Δ, 6,9 + Δ) — доверительным интервалом; Δ вычисляется по формуле

$$\Delta = z\sqrt{\frac{S^2}{n}}, \tag{1.5}$$

где z — критические точки нормального распределения; S^2 — дисперсия анализируемого показателя; n — объем выборки.

Нетрудно видеть, что $\sqrt{\frac{S^2}{n}} = \text{с.о.}\bar{x}$ (см. 1.3).