2. АНАЛИЗ СВЯЗЕЙ МЕЖДУ НОМИНАЛЬНЫМИ ПРИЗНАКАМИ

2.1. Анализ номинальных данных как одна из главных задач социолога

В данном разделе мы коротко покажем, что номинальные данные – главный интересующий социолога вид исходной информации; а анализ связей между признаками – главный вид задач, встречающийся практически в любом эмпирическом социологическом исследовании.

2.1.1. Роль номинальных данных в социологии

Роль номинальных данных в социологии огромна. Объяснить это можно следующими (взаимосвязанными) причинами.

Во-первых, именно номинальные данные чаще всего используются социологами. Вероятно, это можно объяснить сравнительной простотой их получения, естественностью интерпретации, интуитивной уверенностью в состоятельности последней.

Во-вторых, номинальные данные являются более надежными, чем данные, полученные по шкалам более высокого типа, в том смысле, что за ними обычно не стоят трудно проверяемые модели восприятия (имеется в виду восприятие респондентом предлагаемых ему для оценки объектов, суждений, мнений и т.д.; о моделях, предполагаемых известными методами шкалирования см. [Толстова, 1998]), и, в соответствии с этим, при их интерпретации не используются сложные, и зачастую, сомнительные допущения.

В-третьих, в методах, используемых для анализа номинальных данных, обычно бывают «заложены» модели, не вызывающие сомнения, отвечающие естественной логике социолога, изучающего собранную информацию «вручную», без использования математики и ЭВМ. Надеемся, что все сказанное ниже позволит читателю в этом убедиться.

Здесь сделаем небольшое отступление. Среди социологов бытует мнение о том, что достижение интервального уровня измерения всегда является желаемым, поскольку расширяет возможности исследователя, давая ему основания использовать традиционные методы математико-статистического анализа данных. С одной стороны, это, конечно, так: подобные основания действительно имеют под собой почву (хотя надо иметь в виду, что и интервальные данные – не совсем числовые и поэтому к ним применимы не все упомянутые традиционные алгоритмы). Но, с другой стороны, остается вопрос о том, не слишком ли дорога соответствующая цена, не обесценивается ли полученное преимущество несостоятельностью анализируемых данных. Последнее соображение настолько важно, что некоторые авторы вообще полагают, что в социологии только номинальные шкалы имеют право на существование [Чесноков, 1986]. И принять это соображение во внимание имеет смысл еще и потому, что для анализа номинальных данных имеется много достаточно эффективных методов.

2.1.2. Соотношение между причинно-следственными отношениями и формальными методами их изучения

Изучение связей между переменными, как правило, интересует исследователя не само по себе, а как отражение соответствующих причинно-следственных отношений. Представляется излишним доказательство актуальности соответствующих задач, их важность для любого социологического исследования. Однако причинные отношения при изучении социальных явлений не удается выделить в «чистом» виде. Социолог может наблюдать только соответствующие статистические закономерности (статистические

связи), в качестве измерителей которых и выступают известные показатели связи (далее мы увидим, в чем именно проявляется статистичность интересующих нас связей). То устойчивое, необходимое, что скрывается за каждым коэффициентом (или за системой таких коэффициентров) зачастую оказывается возможным отождествить с соответствующей причинной зависимостью.

Подчеркнем, однако, – понятия «причина» и «следствие» в принципе не могут быть формализованы. Никакая математика не может нам доказать, что такой-то признак служит причиной (следствием) того или иного явления. Можно привести массу примеров, когда наличие даже самой сильной статистической связи совершенно не означает наличие соответствующей причинной зависимости. Например, у людей, как правило, одновременно появляется желание надеть легкое платье и пойти искупаться не потому, что одно причинно обусловливает другое, а потому, что оба эти желания вызваны одним и тем же обстоятельством – наступлением жаркой погоды. Другой пример: два студента одновременно вдруг проявляют необыкновенную тягу к знаниям или, напротив, стремятся отлынивать от занятий не потому, что один на другого причинно воздействует, а потому, что сессия у них в одно и то же время – одновременное причинное воздействие третьего признака на каждый из двух данных вызывает статистическую связь между данными признаками. Подобные статистические, не являющиеся причинно-следственными, связи в литературе носят название ложной корреляции. Название не очень удачное – корреляция-то (т.е. статистическая связь) как раз истинна, ложно – причинно-следственное отношение.

Итак, математические методы могут лишь навести нас на мысль о существовании причинных отношений, заставить быть более уверенными в своих предположениях или, напротив, усомниться в них, скорректировать свои априорные представления или даже совсем отказаться от них. Тем не менее, термины «причина» и «следствие» часто употребляются при математическом анализе социологических данных. Однако обычно они отражают лишь априорные исследовательские предположения соответствующего плана.

Правда, в одной из известных ветвей многомерного статистического анализа – так называемом причинном (путевом) анализе [Хейс, 1981] термин «причина» используется именно как нечто формально недоказуемое. В его рамках специально изучаются ситуации с ложными корреляциями, подробно рассматривается, как сложные, опосредованные цепочки причинных отношений могут объяснять их наличие, позволяет понять, за счет чего иногда между какими-то признаками может быть сильная статистическая зависимость при полном отсутствии причинно-следственной, какими сложными опосредованными причинными отношениями эта связь может объясняться.

2.1.3. О понятии таблицы сопряженности

Представляется естественным использовать для оценки связей между признаками так называемые частотные таблицы, или таблицы сопряженности (по существу мы о них уже говорили – это выборочные оценки вероятностных распределений многомерных случайных величин; так, в таблице 3 части I приведен пример распределения для двумерной величины). Заметим, что последний термин обязан своим происхождением именно тому обстоятельству, что на основе анализа подобных таблиц можно судить о сопряженности (совместной встречаемости) каких-то значений одних признаков с некоторыми значениями других признаков. Как мы увидим, связь между номинальными признаками, собственно говоря, и выражается в виде подобных сопряженностей.

Предположим, что мы имеем два признака X и Y, первый из которых принимает «r» значений 1, 2, . . . , r, а второй – «c» значений 1, 2, . . . , c. Назовем двумерной таблицей сопряженности (двумерной частотной таблицей) некоторую матрицу, на пересечении i-й строки и j-го столбца которой стоит число n_{ij} , означающее количество объектов, обладающих i-м значением первого признака и j-м значением второго (i =1, . . . , r; j =1, . . . , c) (использование латинских букв r и c в указанном смысле принято

в литературе; эти буквы сопрягаются с английским словами raw и column, означающими «строка» и «столбец» соответственно; это не позволяет нам забывать, что значения одного признака отвечают строкам таблицы сопряженности, а другого – столбцам). Другими словами, таблица сопряженности выглядит так:

Таблица 6

Общий вид таблицы сопряженности

$$\| n_{ij} \| = \begin{Vmatrix} n_{11} & n_{12} & \dots & n_{1c} \\ n_{21} & n_{22} & \dots & n_{2c} \\ \dots & \dots & \dots & \dots \\ n_{r1} & n_{r2} & \dots & n_{rc} \end{Vmatrix}$$

Обычно ее представляют в несколько ином виде, с явно обозначенными наименованиями признаков и их значений и выписанными маргинальными суммами:

Таблица 7

Общий вид таблицы сопряженности

X	Y						Маргиналы по строкам
	1	2	...	j	...	c	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...	...	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.}$
...	...	...	...	...	...	...	...
r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	$n_{r.}$
Маргиналы по столбцам	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	n

Правый крайний столбец образуют строковые маргинальные суммы (маргиналы по строкам). Величина $n_{i.}$ равна сумме элементов i-й строки (т.е. числу тех объектов, для которых первый признак принимает значение i). Нижняя строка образуется столбцовыми маргинальными суммами (маргиналами по столбцам). Величина $n_{.j}$ равна сумме элементов j-го столбца (т.е. числу тех объектов, для которых второй признак принимает значение j). n – объем выборки, он равен сумме маргиналов по столбцам (либо по строкам).

В последние годы в литературе все чаще используется расширительное понимание таблицы сопряженности. Предполагается, что в качестве ее элементов могут фигурировать не только частоты, но и многие другие числа: скажем, в клетках половозрастной таблицы могут стоять средние значения зарплаты тех людей, которые характеризуются отвечающим клетке значениям пола и возраста. Таким же образом в клетки таблицы могут быть помещены средние другого рода (мода, медиана), дисперсии, величины отклонений от средних по строке (столбцу), разница между эмпирической и теоретической частотой (см. п.2.2.1) и т.д. [Ростовцев и др., 1997, с.177-179]. О том же расширительном понимании таблицы сопряженности говорится в описании известного пакета SPSS.

Ниже, приводя примеры, под объектами, число которых подсчитывается при построении таблицы сопряженности, мы будем иметь в виду респондентов. Хотелось бы, чтобы читатель давал себе отчет в условности таких примеров, понимая, что отнюдь не только респонденты могут интересовать социолога.

2.2. Классификация задач анализа связей номинальных признаков

2.2. 1. Диалектика в понимании признака и его значений

В п.2.3 мы начнем описание ряда методов анализа номинальных данных. Придадим цельность нашему изложению путем установления связи между этими методами посредством прослеживания определенного родства заложенных в этих методах моделей. Сделаем это посредством выработки единого основания для классификации всех рассматриваемых алгоритмов, основания, связанного с определенной типологией социологических задач.

Предлагаемое основание будет опираться на то обстоятельство, что для социолога важно осознание необходимости определенной

диалектики в понимании признака и его значений: выделение ситуаций, когда отдельной альтернативе имеет смысл придать статус самостоятельного признака.

Приведем пример. Нас может интересовать, каким является отвечающее респонденту значение признака «профессия», а может – является ли этот респондент или не является учителем. Во втором случае мы придали статус признака одному значению признака «профессия» – тому, которое называлось «учитель». К такому переходу нас подталкивает не желание пооригинальничать, а стремление адекватно решать стоящие перед социологом задачи. Скажем, изучая связи между рассматриваемыми переменными, мы можем прийти к выводу, что профессия никак не связана с полом (забегая вперед, скажем, что такой вывод можно сделать, использовав какой-либо из известных коэффициентов связи, рассчитывающихся на базе таблицы сопряженности «пол–профессия», скажем, критерий «Хи-квадрат», см. п.2.3.1). Тем не менее, та же статистика может нам говорить, что почти все учителя – женщины, т.е. что соответствующее отдельное значение признака «профессия» связано с полом. Чтобы не «упустить» эту «локальную» связь, мы и должны рассмотреть отдельный дихотомический признак «быть учителем» с целью измерения величины его связи с признаком «пол».

Описанное требование можно обобщить: самостоятельной переменной может отвечать не одно значение некоторого признака, а сочетание таких значений (скажем, при решении ряда задач имеет смысл объединить, учителей и врачей вместе), каждое из которых соответствует, вообще говоря, своему признаку (о таких ситуациях, когда объединяются альтернативы разных признаков, пойдет речь в п.2.5).

Два слова о терминах. В работе [Чесноков, 1982] предлагается называть *глобальными* коэффициенты парной связи, рассчитывающиеся на основе учета всех градаций рассматриваемых признаков, и *локальными* – коэффициенты связи, рассчитывающиеся на основе учета одной градации одного признака и одной градации другого. Нам представляется неприемлемым деление всех показателей на глобальные и локальные, поскольку при таком

подходе из рассмотрения (во всяком случае на терминологическом уровне), выпадают связи «промежуточных» видов: такие, когда учитываются несколько градаций каждого признака. Однако термин «локальная связь» мы будем использовать, понимая под таковой связь между отдельными альтернативами.

Заметим, что приведенные выше соображения имеют самое непосредственное отношение к проблеме социологического измерения, к анализу понятия «признак» и, в конечном счете, к проблеме операционализации понятий, к изучению перехода от реальных многогранных объектов к их узкому, всегда ограниченному описанию набором некоторых признаков (к «мышлению признаками», по выражению автора работы [Ноэль, 1993]).

Описанные ситуации возникают в силу того, что, с одной стороны, само понятие признака имеет смысл только при некоторой однокачественности тех объектов, для которых значения признаков вычисляются; с другой стороны, – каждому значению признака отвечает свое собственное качество. Понятие однокачественности относительно. На разных этапах исследования может возникнуть потребность однокачественные объекты считать разнокачественными и наоборот. Так, выше мы показали, что бывают ситуации, когда однокачественными объектами мы считаем всех тех и только тех респондентов, которые имеют профессию учителя. Человек же с профессией врача в такой ситуации будет иметь другое качество. При изучении проблем интеллигенции учитель и врач могут стать однокачественными объектами. Если же мы работаем с признаком «профессия» как единым целым, то тем самым полагаем, что этот признак отражает существование некоторого социального института и однокачественными являются все члены такого общества, в котором этот институт имеется.

В обосновании необходимости «склеивания» отдельных значений разных (вообще говоря) признаков просматривается актуальность решения следующей проблемы социологического измерения: чтобы отразить латентные свойства объекта, мы вынуждены «выдергивать» отдельные значения разных признаков, формировать из этих «надерганных» значений различные

комбинации, надеясь, что какое-то сочетание хотя бы частично явится индикатором определенного «поведения» объекта.

Дальнейшее обобщение требования склеивания отдельных градаций приводит к осознанию возможности рассмотрения в качестве нового признака не сочетания отдельных альтернатив, а сочетания нескольких признаков. Соответствующее обобщение проблемы измерения очевидно: новым измеряемым признаком является здесь комбинация исходных признаков.

Продолжая ту же логику, естественно приходим к необходимости рассмотрения всех признаков сразу как единой системы.

Выделение перечисленных возможностей мы будем рассматривать как основу для дальнейшего изложения (в частности, для классификации методов анализа связей номинальных признаков).

Итак, в соответствии с предлагаемой точкой зрения, каждый рассматриваемый метод можно трактовать как реализацию следующего процесса: все исходные номинальные признаки как бы «рассыпаются» на отдельные градации, которые затем по-разному комбинируются, на их основе строятся новые признаки, взаимоотношения которых далее изучаются. Каждый метод анализа связей номинальных данных предлагается рассматривать как метод поиска либо связей между разными группами альтернатив, либо групп альтернатив, определяющих некоторое поведение респондентов (задаваемое разными способами). Методы систематизируются в зависимости от отвечающих им способов агрегирования отдельных альтернатив в новые признаки.

Использование предлагаемого подхода, на наш взгляд, побуждает исследователя не забывать о существовании многих методов, весьма адекватных социологическим задачам, но мало используемых социологами.

В данном разделе мы будем рассматривать методы, которые включаются в указанную классификацию. Но прежде, чем более подробно ее описать (что будет сделано в п.2.2.2), представляется важным рассмотреть один момент, позволяющий лучше понять, как модели, заложенные в интересующих нас методах, соотносятся с моделями других известных методов анализа данных (о других

моментах такого рода см. п.2.2.3).

Нетрудно заметить, что упомянутые выше задачи (и отвечающие им методы), связанные с поиском групп альтернатив, определяющих некоторое поведение респондентов, очень похожи на задачи поиска того, что в математической статистике (в частности, в дисперсионном и регрессионном анализах; описание первого можно найти, например, в [Статистические методы..., 1979], о втором пойдет речь в п.2.6) называется *взаимодействием*.

Напомним, что использование этого термина предполагает выделение среди всех признаков детерминируемого, зависимого и группы детерминирующих его, независимых признаков (подробнее о подобных терминах см. п.2.5.3.1). «Взаимодействие» означает сочетание значений независимых признаков, определяющих тот или иной уровень зависимого (заметим, что в дисперсионном анализе зависимый признак предполагается количественным, т.е. таким, значения которого получены, по крайней мере, по интервальной шкале; а совокупность независимых признаков фиксируется). Например, при изучении миграционного поведения взаимодействием может служить свойство респондента одновременно быть мужчиной (т.е. обладать, скажем, значением «1» признака 4 – «пол») и иметь высшее образование (т.е. обладать, например, значением «5» признака 6 – «образование»), если это свойство детерминирует желание обладающего им человека уехать за границу.

Роль поиска взаимодействий в эмпирической социологии вряд ли можно преувеличить. Однако представляется, что потребность практики делает целесообразным расширение этого понятия. Для того, чтобы пояснить, каким способом это можно сделать, попытаемся вдуматься в смысл того, что значит делать какие-то выводы в терминах рассматриваемых (номинальных) признаков. Вероятно, исходя из здравого смысла, подобные выводы должны иметь вид (мы имеем в виду формальную структуру того статистического утверждения, которое служит социологу основой для дальнейших выводов о причинно-следственных отношениях):

« 5-е значение 8-го признака часто встречается с 3-м значением 14-го и 1-м значением 2-го» , «из того, что 3-й признак принимает

2-е значение одновременно с тем, что 4-й принимает 5-е значение, как правило, следует, что 6-й признак принимает либо 2-е, либо 3-е», «из того, что 3-й признак принимает какое-либо значение, кроме 2-го, следует, что 7-й признак принимает 4-е значение» и т.д. (надеемся, что для понимания сказанного не требуется более конкретно формулировать подобные утверждения: скажем, указывать, что 3-й признак – это возраст, его 5-е значение – указание того, что возраст конкретного респондента заключен в интервале от 35 до 40 лет и т.д.).

(Выражения, подобные сформулированным, являются наиболее естественными для социолога. Они отвечают сути номинальных шкал, тому, что каждое значение признака означает самостоятельное автономное качество объекта. Однако исследователь зачастую стремится по-другому формулировать искомые содержательные выводы, вольно или невольно вписывая их в традиционные рамки классических математико-статистических формулировок: «такие-то два признака имеют сильную статистическую связь», «второй признак линейно зависит от седьмого» и т.д. Можно показать, что такие формулировки тоже могут быть «переведены» на язык наших взаимодействий.)

Анализ подобного рода выражений заставляет следующим образом *обобщить понятие взаимодействия*:

- совокупность признаков-предикторов будем считать «плавающей» (естественно, – в пределах множества признаков, заданных в исследовании; напомним, что в дисперсионном анализе фиксируется небольшое количество признаков-предикторов и рассматриваются все возможные сочетания их значений; среди этих значений и ищутся взаимодействия); в частности, будем полагать, что какое-то сочетание значений одного набора предикторов может определять одно значение признака-функции, а некоторое сочетание значений другого набора предикторов – другое значение функции; например, в добавление к высказанному выше гипотетическому предположению о том, что у мужчин с высшим образованием появляется желание покинуть Родину, можно добавить еще одно предположение – о том, что женщины, имеющие более двух детей, напротив, выступают против отъезда за границу;

- будем полагать, что взаимодействием может быть не только конъюнкция суждений типа «значение такого-то признака равно тому-то» (именно конъюнкцией суждений «человек – мужчина» и «человек имеет высшее образование» является суждение «человек является мужчиной с высшим образованием»), а любые логические функции от таких выражений (предполагаем, что читатель знает определение основных логических функций – конъюнкции, дизъюнкции, импликации, отрицания; используемые здесь и ниже сведения по логике можно почерпнуть, например, из [Бочаров, Маркин, 1994]); например, взаимодействием будем считать суждение «человек является или пенсионером, или женщиной с маленьким ребенком, или не бизнесменом», если люди, обладающие соответствующими свойствами, не желают покидать родные места; (сравним также с упомянутыми выше «2-м значением 3-го признака и 5-м – 4-го, любым значением 3-го, кроме 2-го»;) такого рода функции будем называть *объясняющими, или детерминирующими, положениями (выражениями);* их будем описывать так, как это обычно делается в литературе: используя для обозначения входящих в них признаков букву X с индексами ($X_3(2)$ & $X_4(5)$, $\neg X_3(2)$ и т.д.).
- будем полагать, что наше взаимодействие может определять не только некоторое значение непрерывного признака (как в дисперсионном анализе), но и любую логическую функцию значений произвольных, в том числе дискретных (в частности, номинальных) признаков (ср. упомянутые выше «3-е значение 14-го признака и 1-е – 2-го; 2-е или 3-е значение 4-го признака); каким-либо другим образом задаваемое «поведение» респондента (примеры будут приведены в п.2.5, при обсуждении алгоритмов THAID и CHAID); частоту в таблице сопряженности (ср. «5-е значение 8-го признака часто встречается с 3-м значением 14-го и 1-м значением 2-го»; это мы рассматривать не будем; однако подчеркнем, что речь идет об очень актуальных для социологии задачах, решаемых с помощью логлинейного анализа [Аптон, 1982]); а может и ничего не определять, но тогда естественно требовать просто истинность взаимодействия как логической функции; то, что определяет взаимодействие, будем называть

объясняемыми, или детерминируемыми, положениями; их будем описывать обычно, используя для входящих в них признаков букву Y с индексами.

О поиске обобщенных взаимодействий будем говорить как о поиске закономерностей или детерминаций.

Рассмотрим еще одну сторону понимания термина «взаимодействие» – то, каким образом могут быть связаны объясняющее и объясняемое положения. Обратим внимание на некоторые аспекты приведенных выше формулировок типичных социологических утверждений в терминах используемых номинальных признаков. «5-е значение 8-го признака **часто встречается** с 3-м значением 14-го и 1-м значением 2-го», «из того, что 3-й признак принимает 2-е значение одновременно с тем, что 4-й принимает 5-е значение, **как правило**, следует, что 6-й признак принимает либо 2-е, либо 3-е». Представляются очевидными причины появления выделенных слов в приведенных выражениях. Мы имеем дело лишь со статистическими закономерностями, являющимися в определенном смысле приближенными. Например, если даже вполне можно считать, что мужчины с высшим образованием имеют склонность эмигрировать, практически всегда из этого правила будут исключения. И всегда встает вопрос о том, каково должно быть количество подобных исключений для того, чтобы мы все-таки считали найденную закономерность закономерностью. К этому вопросу мы не раз будем возвращаться.

Как формализовать выражения «часто встречается», «как правило» и т.д.? Без формализации мы не можем проверять справедливость рассматриваемых суждений. Формализация же – это фрагмент используемой модели. Он разный в разных методах. Так, в неоднократно упомянутом нами дисперсионном анализе речь идет о статистической значимости различий средних значений выходного признака для респондентов, обладающих разными сочетаниями значений предикторов. Как мы увидим ниже, в других интересующих нас алгоритмах задействованы другие критерии (о них пойдет речь ниже, при описании соответствующих алгоритмов). Возможность разных критериев тоже может рассматриваться как элемент обобщенного подхода к пониманию

взаимодействия. Обсуждая подобные критерии, будем говорить о формализации *понятия приближенности* искомой закономерности.

При таком понимании взаимодействия можно сказать, что поиск взаимодействий разного рода служит основой большинства рассматриваемых нами методов анализа номинальных данных. В следующем параграфе будут приведены примеры.

2.2.2. Классификация рассматриваемых задач и отвечающих им методов

Ниже в скобках бы будем указывать примеры математических методов, направленных на решение задач выделяемых классов. При первом чтении это можно опустить. Мы называем конкретные методы уже сейчас, до того как они будут описаны (а следующие параграфы будут посвящены такому описанию; сами названия этих параграфов отвечают названиям выделенных ниже классов задач), по двум причинам: во-первых, для того, чтобы читатель, знакомый с упоминаемыми методами, лучше понял нашу классификацию; во-вторых, мы надеемся,что читатель вернется к настоящему параграфу после прочтения всей книги с целью более четко представить себе совокупность тех алгоритмов, из числа которых ему предстоить выбрать инструмент для обнаружения интересующих его закономерностей.

Итак, в соответствии с предлагаемым основанием выделяются задачи типа:

– «альтернатива–альтернатива», т.е. такие, которые позволяют изучать связь между отдельными значениями любых рассматриваемых признаков (примером является детерминационный анализ [Чесноков, 1982]);

– «(группа альтернатив) – (группа альтернатив)» (анализ фрагментов таблиц сопряженности [Интерпретация и анализ..., 1987, гл.2], алгоритмы типа «пятна» и «полосы» [Ростовцев, 1985, с.203-214]); эту группу методов можно расширить, условно назвав результат такого расширения методами типа:

– «(группа альтернатив) – («поведение» объектов)», где «поведение» (подчеркнем, – не одного объекта, а целой сово-

купности, заданной рассматриваемой группой альтернатив; такое «поведение» в определенном смысле есть описание этой совокупности, которое, в свою очередь, можно интерпретировать как характеристику некоторого типа объектов) может пониматься по-разному: как определенный каким-либо образом «средний» уровень заранее заданного результирующего признака (скажем мы можем искать тип людей с низким уровнем зарплаты и тип людей с высоким уровнем зарплаты), как истинность для рассматриваемой совокупности некоторой логической функции от элементарных формул типа P(a) = 1 (так называемых логических закономерностей), где буквой P обозначен произвольный признак, а приведенное выражение означает: «значение признака P для объекта a равно 1» и т.д. (методы выявления логических закономерностей [Лбов, 1981], методы поиска детерминирующих сочетаний значений рассматриваемых признаков, в том числе известные на Западе алгоритмы, для обозначения которых используются аббревиатуры, включающие в себя сочетание AID (automatic interaction detector): THAID [Интерпретация и анализ..., 1987, с.136-151; Messenger, Mandell, 1972; Morgan, Messenger, 1973]), CHAID [Agresti, 1990; Magidson, 1993; Derrick, Magidson, 1992], AID3 [Sonquist, Morgan, 1973] и т.д. Сравнение THAID и AID3 осуществляется в [Kass, 1980]. Ряд методов описан в [Типология и классификация..., 1982, с.213-231]. Назовем также брошюру [Ливанова, 1990], где подробно описан процесс реализации на ЕС ЭВМ алгоритма AID3. Хотя в наше время персональных компьютеров такое описание не является актуальным, тем не менее, на наш взгляд, указанная работа не стала бесполезной для социолога, поскольку в ней помимо правил обращения с ЭВМ серии ЕС подробно раскрывается сущность самого алгоритма).

Частным случаем упомянутых комбинаций явится объединение в одну группу альтернатив, отвечающих одному признаку. В соответствии с этим, выделим класс задач:

– «признак–признак» (традиционные, наиболее знакомые социологу коэффициенты парной связи).

Продолжая рассуждения, отвечающие той же логике, нетрудно прийти к выводу, что та же специфика измерительных процедур

может вызвать потребность объединять не только «надерганные» из разных признаков альтернативы, но и признаки в целом. В соответствии с этим, в рамках нашей классификации выделим группы методов:

– «признак – (группа признаков)» (регрессионный анализ, многие методы построения индексов);

(Отметим, что при использовании регрессионного анализа зачастую решаются также задачи типа «(группа альтернатив) – («поведение» объекта)»; это ярко демонстрирует его так называемый номинальный вариант [Аргунова, 1990; Типология и классификация..., 1982; Hardy, 1993], см. также п.2.6.)

– «(группа признаков) – (группа признаков)» (канонический анализ [Интерпретация и анализ..., 1987]). Это известный математико-статистический метод. Однако он крайне редко используется социологами, считающими его типично «количественным» методом. В действительности же соответствующий подход является актуальным для анализа именно номинальных данных: он дает возможность осуществлять их оцифровку (т.е. приписать каждому значению номинального признака некоторое число), изучать связи между признаками с так называемыми «совместными» альтернативами, эффективно находить веса признаков при формировании из них индекса. Идеи, заложенные в каноническом анализе используются в таком широко применяющимся в современной западной социологии (в том числе в ставших «модными» в России маркетинговых исследованиях) методе, как корреспонденс-анализ, или анализ соответствий [Clausen, 1998]).

Тип задач, отвечающих рассмотрению всей совокупности признаков как системы, назовем так:

– анализ системы признаков (логлинейный анализ [Аптон, 1982; Елисеева, Рукавишников, 1977; Мирзоев, 1980, 1981; Миркин, 1980]; причинный анализ [Елисеева, Рукавишников, 1982; Осипов, Андреев, 1977; Хейс, 1981]).

К сожалению, в настоящей работе мы не имеем возможности рассмотреть последние два типа задач.

Конечно, если строго следовать формальной логике, можно заметить, что почти все упомянутые классы методов могут быть

сведены к одному – классу «(группа альтернатив) – (группа альтернатив)», поскольку с формальной точки зрения частным случаем группы альтернатив является и отдельная альтернатива; и набор градаций, отвечающих одному признаку; и совокупности значений сразу нескольких признаков. Но с содержательной точки зрения все же мы не можем игнорировать различие между выделенными выше совокупностями альтернатив. В частности, понятие признака – это нечто, отвечающее вполне определенной социальной реальности. За частью альтернатив признака эта реальность не стоит. И, как мы увидим ниже, методы, позволяющие решать задачи выделенных классов, различны, поскольку различны постановки соответствующих содержательных вопросов.

Казалось бы, изложение надо начинать с описания наиболее простых методов – типа «альтернатива – альтернатива». Однако исторически сложилось так, что сначала были разработаны коэффициенты парной связи между признаками (т.е. наши методы типа «признак – признак»). А все остальные подходы опирались на соответствующие теоретические положения. Мы не хотим претендовать на разработку новых подходов к обоснованию известных коэффициентов. Поэтому начнем как бы с середины нашей схемы – с описания методов измерения связей между двумя номинальными признаками. Однако прежде позволим себе некоторое отступление от основного содержания настоящей книги. Дело в том, что подходами, рассматриваемыми в настоящей работе, отнюдь не ограничивается ни совокупность всех методов анализа номинальных данных вообще, ни совокупность методов анализа связей между номинальными переменными. Для того, чтобы более четко охарактеризовать круг задач, решение которых становится доступным с помощью подходов, описанных в следующих параграфах, попытаемся очертить то место, которое эти подходы занимают в гораздо более широкой совокупности известных методов анализа номинальных данных. Сделаем это, обратившись к рассуждениям, нетрадиционным для работ по анализу данных.

2.2.3. Выделение двух основных групп методов анализа номинальных данных. Место рассматриваемых подходов в этой группировке

Специфичность настоящего параграфа состоит в том, что мы попытаемся достичь сформулированной цели с помощью установления связи между идеями математики и теоретической социологии. Говоря подробнее, мы на примере покажем, что математик зачастую ставит перед собой те же вопросы, что и социолог, но специфика ответов у каждого специалиста (понятия «математик» и «социолог» мы здесь интерпретируем как некоторые идеальные типы, как отражение разницы видения мира разными исследователями, разницы, обусловленной различием их природных данных, склада ума, той среды, в которой они формировались как ученые и т.д.) своя.

«Математик» в большей мере умеет вычленить в реальности какие-то поддающиеся формализации, строгому описанию фрагменты. При этом может не только использовать известный математический язык, но и создавать новый (достаточно формализованное, строгое описание каких-то аспектов реальности, по определению, называется математическим). Ясно, что строгость описания реальности сопряжена со сравнительной ограниченностью, бедностью описываемого. «Социолог» дает более расплывчатое описание увиденного. Но расплывчатость эта зачастую обусловливается более широким кругозором, пониманием того, что отнюдь не все важные для социологии аспекты реальности поддаются формализации, по крайней мере, при современном развитии науки (в свете сказанного представляется очевидной причина того, почему Конт в своей известной классификации наук самой простой наукой назвал математику, а самой сложной – социологию).

Два слова о том, почему мы сочли нужным включить в книгу настоящий параграф. Задуматься о глубинных связях социологии и математики автора побудила необходимость решить известную проблему преподавания студентам-социологам дисциплин,

связанных с использованием математического аппарата. Как мы уже отмечали, студенты часто отторгают такие дисциплины, полагая, что они являются чужеродными для социолога. «Противоядием» против такого отторжения обычно служит демонстрация студентам многочисленных примеров использования в эмпирической социологии методов анализа данных (либо методов математического моделирования разного рода социальных явлений и процессов). «Хорошие» студенты начинают понимать, что математика необходима им для будущей практической работы с эмпирическими данными. Однако при этом никакой глубинной связи между социологией и математикой не усматривается. Само собой разумеющимися обычно считаются следующие положения.

1) Да, математика помогает социологу охватить единым взором огромные массивы, коротко выразить суть содержащихся в них статистических закономерностей, взаимосвязей между отдельными явлениями и т.д. 2) Но к получению наиболее интересных для социолога фактов эмпирической социологии, связанных с серьезным анализом причинно-следственных отношений математика имеет слабое отношение, поскольку она использует методы, разработанные в основном для естественных наук, и поэтому позволяет улавливать зависимости, хотя и важные для социолога, но не носящие специфически социологического характера. 3) Более того, к поиску закономерностей, касающихся глубокого анализа сознания респондента, математика вообще не имеет отношения. Этот более глубокий анализ связывается обычно с пониманием, а не с объяснением. Соответствующее знание можно получить только с помощью так называемых качественных методов. 4) Тем более математика далека от того, с чем имеет дело так называемая теоретическая социология.

Определенные размышления позволили нам прийти к несогласию с положениями (2), (3), (4). На наш взгляд, связь между математикой и социологией гораздо глубже, чем это принято считать. То, что студенты ее не видят, представляется естественным. Изучением такой связи наша наука практически не занималась. Лишь в самые последние годы в работах специалистов

по теоретической социологии стали появляться параграфы с названиями: «Программа статистически-вероятностно ориентированной науки об обществе» (о творчестве Кондорсе), «Идея инкорпорирования учения о социальном прогрессе в математическое естествознание» (о творчестве И.Канта) [Давыдов, 1995]. Однако соответствующий контекст наводит на мысль о том, что эти словосочетания отражают скорее некие интуитивные догадки, пожелания на будущее, чем конструктивный подход к изучению общественных закономерностей с помощью математического аппарата. Ниже мы по существу попытаемся внести некоторый элемент конструктивности в понимание связи идей математики и теоретической социологии.

Перейдем к выделению интересующих нас групп методов.

Во Введении мы уже предложили некоторую группировку (классификацию) методов анализа данных – деление их на методы дескриптивной статистики, анализа связей между признаками, классификации объектов и поиска латентных переменных. Однако эта классификация является довольно грубой, носит весьма относительный характер и в весьма слабой мере опирается на более или менее серьезные (с точки зрения глубинных моментов, мешающих адекватности использования математики в социологии) модельные предпосылки.

Выделим в огромной совокупности методов анализа номинальных данных два мощных направления, стихийно сложившихся в мировой науке. За каждым из них стоит своя методологическая концепция, свой круг решаемых задач. Глубинные методологические предпосылки, лежащие в основании такого выделения, касаются рефлексии социолога по поводу процесса формирования используемых в исследовании понятий, связаны, в частности, с известным многовековым обсуждением вопросов о номинализме и реализме в социологии. Напомним, о чем идет речь.

Начало упомянутых рассмотрений относится к известному спору об «универсалиях» средневековых схоластов (спор об отношении общего к единичному) [Краткий очерк..., 1960, с.111]. «Реалисты» полагали, что «универсалии» (общие роды) существуют реально, независимо от человеческой мысли и речи.

«Номиналисты» – что «универсалии» не существуют реально, независимо от человека. Они суть только общие имена (например, «человек вообще», как родовая общность, не существует; реально существуют только отдельные люди; «человек» – лишь общее имя, которым называется каждый конкретный человек).

Среди авторов методов анализа данных также можно выделить своеобразных «реалистов» и «номиналистов». И показать это можно, обратившись к анализу выделяемых нами направлений.

Предлагаемая классификация опирается на некоторые фундаментальные модельные предположения о характере используемых номинальных признаков. Имеется в виду возможность различной интерпретации номинальных данных. Речь идет о том, считаем ли мы, что значения каждого номинального признака являются самостоятельными сущностями, отвечающими разным качествам изучаемых объектов (что часто отождествляется с «превращением» каждого значения в автономный дихотомический признак; о такой дихотомизации пойдет речь в п.2.6.3), или же полагаем, что за этими значениями (сочетаниями таких значений) стоит некоторая непрерывная (случайная) величина. В последнем случае мы опираемся на предположение о том, что номинальность наблюдаемого признака объясняется нашим неумением точно измерить «стоящую» за признаком переменную (заметим, что здесь мы не касаемся затронутой выше проблемы, связанной с возможностью рассмотрения каждого найденного с помощью некоторых приемов анализа данных сочетания значений каких-либо признаков как значения строящегося одномерного индекса, см. начало п.2.2.1).

Так, можно рассматривать профессию как единое целое, а можно отдельно рассмотреть свойство «Быть учителем», или свойство «Иметь профессию, представителей которой относят к интеллигенции» и т.д.

Выделение указанных подходов к интерпретации номинальных данных представляется достаточно принципиальным по крайней мере по двум причинам.

Первую причину можно назвать *гносеологической*. Именно анализируя возможность усматривать за наблюдаемым признаками некоторую скрытую непрерывную переменную, мы попадаем в самую гущу интересующего нас спора между сторонниками социологического реализма и социологического номинализма Если мы полагаем, что отдельные градации какого-либо признака представляют собой самостоятельные сущности, т.е. отказываемся пользоваться предположением о существовании некоторой переменной, стоящей за ними, то тем самым встаем на сторону номинализма. В таком случае мы полагаем, например, что существуют люди–учителя, люди–токари, а вот понятие «профессия человека» – это лишь некоторое введенное для удобства и лишенное всякого онтологического содержания название совокупности людей, рассматриваемых как носителей указанных свойств. В такой ситуации столь же бессодержательной будет фраза: «пол и профессия статистически связаны друг с другом». Но вполне осмыслено высказывание: «почти все учителя – женщины».

Если же мы считаем, что наблюдаемые значения – это лишь разные проявления некоторой объективно существующей непрерывной латентной переменной, т.е. некоторого общего для всех людей (системного) качества, то тем самым переходим на позиции социологического реализма (во всяком случае, относительно рассматриваемых качеств отдельных людей).

Представляется возможным также связать первую интерпретацию с гуманитарным подходом к измерению, а вторую – с естественно-научным подходом (об этих подходах см. [Чесноков, 1986]; теория гуманитарных измерений принимает как фундаментальный факт способность людей различать образы и поименовывать их).

Таким образом, мы видим, что одна из актуальных для социологии проблем своеобразно, в каком-то узком своем аспекте, рассматривается математикой.

Вторая причина выделения названных подходов к интерпретации номинальных данных – чисто *практическая*. Разные интерпретации приводят к возможности постановки разных задач и, соответственно, – к возникновению (и использованию) разных методов анализа данных.

Первая интерпретация обусловливает то, что во главу угла исследователь ставит поиск сочетаний значений признаков, детерминирующих «поведение» (по-разному понимаемое) респондента, т.е. поиск взаимодействий. Соответствующим методам мы уделим большое внимание.

При второй интерпретации действия исследователя, как правило, бывают направлены на то, чтобы «вытащить» из исходной информации «стоящую за кадром» латентную переменную, найти «истинное» ее значение для каждого респондента. Часто при этом используются идеи так называемой «оцифровки», т.е. приписывания каждой градации любого номинального (порядкового) признака определенного числа, отвечающего искомому «истинному» значению соответствующей латентной переменной. Речь идет о широком круге родственных друг другу статистических методов, активно применяющихся в западной социологии (особенно во Франции, где совокупность этих методов зачастую отождествляется с методами анализа данных), но слабо известных российским социологам. Это анализ соответствий [Адамов, 1991; Дидэ, 1985; Жамбю, 1988; Клишина, 1991; Benzecri, 1973; Clausen, 1998], канонический анализ [Интерпретация и анализ..., 1987; Thompson, 1984], конджойнт-анализ [Louvier, 1988], латентно-структурный анализ (ссылки см. в сноске 6 к части I), собственно алгоритмы оцифровки [Интерпретация и анализ..., 1987, гл.3; Айвазян и др., 1983] и т.д. Сюда же с определенной оговоркой можно отнести методы многомерного шкалирования [Интерпретация и анализ..., 1987, гл.8; Клигер и др., 1978, гл.4; Kruscal, Wish, 1978].

Эти методы, как известно, работают не с матрицами типа «объект–признак», а с матрицами близостей между шкалируемыми объектами; но интересующее нас положение остается в силе: предполагается, что респондент, так или иначе дающий оценку объектам, мыслит последние как точки в некотором пространстве восприятия, оси которого – непрерывные числовые переменные; задача же состоит в нахождении этих переменных (т.е. в определении того, какова их суть, каковы их значения для каждого респондента). Сюда же можно отнести и многие известные методы

построения социологических индексов, например, известные способы одномерного шкалирования, связываемые обычно с именами Терстоуна, Лайкерта, Гуттмана. Перечисленные методы нами рассматриваться не будут.

Однако в рамках второго подхода находятся и некоторые методы другого рода, в том числе методы, позволяющие искать взаимодействия (CHAID) и измерять связь как между номинальными признаками в целом (Хи-квадрат), так и между отдельными группами альтернатив, отвечающих таким признакам (анализ фрагментов таблицы сопряженности). Эти методы будут подробно рассмотрены ниже, а CHAID будет сравнен с теми методами поиска взаимодействий, которые не опираются на существование упомянутой латентной переменной.

2.3. Анализ связей типа «признак–признак»

Для измерения связи между двумя номинальными признаками в литературе предлагается более сотни коэффициентов. Это является следствием того, что интересующее нас явление – указанную связь (еще раз подчеркнем, что мы говорим о статистической связи, хотя в действительности нас, как правило, интересуют соответствующие причинно-следственные отношения) – оказывается возможным формализовать по-разному. И каждому способу формализации отвечает свое понимание сути искомой связи, своя априорная модель того, что мы хотим изучить.

Мы не будем описывать все известные из литературы коэффициенты рассматриваемого характера. Коснемся лишь трех подходов к измерению парной связи между номинальными признаками. Эти подходы являются наиболее употребительными на практике. Надеемся, что их анализ, осуществленный ниже, заставит читателя «почувствовать» ту сложность социальной реальности, которая обусловливает возможность выделения в ней разных сторон, каждая из которых по-своему «представляет» изучаемое явление, по-своему формализуется.

2.3.1. Коэффициенты связи, основанные на критерии «Хи-квадрат»

2.3.1.1. Понимание отсутствия связи между признаками как их статистической независимости

Приведем простой пример, иллюстрирующий рассматриваемый подход к пониманию связи между двумя номинальными признаками. Предположим, что перед нами стоит задача оценки того, зависит ли профессия респондента от его пола. Пусть наша анкета содержит соответствующие вопросы и в ней перечисляются пять вариантов профессий, закодированных цифрами от 1 до 5; для обозначения же мужчин и женщин используются коды 1 и 2 соответственно. Для краткости обозначим первый признак (т.е. признак, отвечающий вопросу о профессии респондента) через Y, а второй (отвечающий полу) – через X. Итак, наша задача состоит в том, чтобы определить, зависит ли Y от X.

Предположим, что исходная таблица сопряженности, вычисленная для каких-то 100 респондентов имеет вид:

Таблица 8

Пример таблицы сопряженности для двух независимых признаков

Профессия	Пол		Итого
	1	2	
1	18	2	20
2	18	2	20
3	45	5	50
4	0	0	0
5	9	1	10
Итого	90	10	100

Вероятно, любой человек согласится, что в таком случае признаки можно считать независимыми, поскольку и мужчины, и женщины в равной степени выбирают ту или иную профессию: первая и вторая профессии пользуются одинаковой популярностью и у тех, и у других; третью – выбирает половина мужчин, но и половина женщин; четвертую не любят ни те, ни другие и т.д. Итак, мы делаем вывод: независимость признаков означает пропорциональность столбцов (строк; с помощью несложных арифметических выкладок можно показать, что пропорциональность столбцов эквивалентна пропорциональности строк) исходной частотной таблицы. Заметим, что в случае пропорциональности «внутренних» столбцов таблицы сопряженности, эти столбцы будут пропорциональны также и столбцу маргинальных сумм по строкам. То же – и для случая пропорциональности строк, они будут пропорциональны и строке маригинальных сумм по столбцам.

Приведенная частотная таблица, полученная эмпирическим путем, является результатом изучения выборочной совокупности респондентов. Вспомним, что в действительности нас интересует не выборка, а генеральная совокупность. Из математической статистики мы знаем, что выборочные данные никогда стопроцентно не отвечают «генеральным». Любая, самая хорошая выборка всегда будет отражать генеральную совокупность лишь с некоторым приближением, любая закономерность будет содержать так называемую выборочную ошибку, случайную погрешность. Учитывая это, мы, вероятно, будем полагать, что, если столбцы выборочной таблицы сопряженности мало отличаются от пропорциональных, то такое отличие скорее всего объясняется именно выборочной погрешностью и вряд ли говорит о том, что в генеральной совокупности наши признаки связаны. Так мы проинтерпретируем, например, таблицу 9 (по сравнению с таблицей 8 в ней четыре частоты изменены на единицу) и, наверное,

Таблица 9

Первый пример таблицы сопряженности, частоты которой мало отличаются от ситуации независимости признаков

Профессия	Пол		Итого
	1	2	
1	17	3	20
2	19	1	20
3	45	5	50
4	0	0	0
5	9	1	10
Итого	90	10	100

Таблица 10

Второй пример таблицы сопряженности, частоты которой мало отличаются от ситуации независимости признаков

Профессия	Пол		Итого
	1	2	
1	16	4	20
2	20	0	20
3	45	5	50
4	0	0	0
5	9	1	10
Итого	90	10	100

Таблица 11

Пример таблицы сопряженности, частоты которой значительно отличаются от ситуации независимости признаков

Профессия	Пол		Итого
	1	2	
1	15	5	20
2	20	0	20
3	46	4	50
4	0	0	0
5	9	1	10
Итого	90	10	100

таблицу 10 (те же частоты изменены на две единицы). А как быть с таблицей 11?

Общая идея здесь ясна: сильное отклонение от пропорциональности заставляет нас сомневаться в отсутствии связи в генеральной совокупности, слабое отклонение говорит о том, что наша выборка не дает нам оснований для таких сомнений. Но насколько сильным должно быть указанное отклонение для того, чтобы описанные сомнения возникли?

Наука не дает точного ответа. Она предлагает нам лишь такой его вариант, который формулируется в вероятностных терминах. Этот ответ можно найти в математической статистике. Чтобы его воспринять, необходимо взглянуть на изучаемую связь, опираясь на своеобразное математико-статистическое видение мира. Опишем соответствующие рассуждения в следующем параграфе.

Сразу скажем, что эти рассуждения типичны для математической статистики – речь идет об одной из основных решаемых ей задач – проверке статистической гипотезы.

2.3.1.2. Статистика «Хи-квадрат» и проверка на ее основе гипотезы об отсутствии связи

Предположим, что мы имеем две номинальных переменных, отвечающую им частотную таблицу типа 7 и хотим на основе ее анализа определить, имеется ли связь между переменными. Будем искать ответ на этот вопрос с помощью проверки статистической гипотезы о независимости признаков. Используя терминологию математической статистики, можно сказать, что речь пойдет о проверке нуль-гипотезы H_0: «связь между рассматриваемыми переменными отсутствует».

Далеко не для каждой интересующей социолога гипотезы математическая статистика предоставляет возможность ее проверки, не для каждой гипотезы разработана соответствующая теория. Но если упомянутая возможность существует, то соответствующая логика рассуждений сводится к следующему.

Допустим, что для какой-то статистической гипотезы H_0 разработана упомянутая теория и мы хотим эту гипотезу проверить. Математическая статистика предлагает некий критерий. Он представляет собой определенную числовую функцию f от наблюдаемых величин, например, рассчитанную на основе частот выборочной таблицы сопряженности: $f = f(n_{ij})$. Представим теперь, что в нашем распоряжении имеется много выборок, для каждой из которых мы можем вычислить значение этой функции. Распределение таких значений в предположении, что проверяемая гипотеза справедлива (для генеральной совокупности), хорошо изучено, т.е. известно, какова вероятность попадания каждого значения в любой интервал. Грубо говоря, это означает, что, если H_0 справедлива, то для каждого полученного для конкретной выборки значения f можно сказать, какова та вероятность, с которой мы могли на него «наткнуться». Вычисляем значение

$f_{выб}$ критерия f для нашей единственной выборки. Находим вероятность $P(f_{выб})$ этого значения.

Далее вступает в силу своеобразный принцип невозможности маловероятных событий: мы полагаем, что если вероятность какого-либо события очень мала, то это событие практически не может произойти. И если мы все же такое маловероятное событие встретили, то делаем из этого вывод, что вероятность определялась нами неправильно, что в действительности встреченное событие не маловероятно.

Наше событие состоит в том, что критерий принял то или иное значение. Если вероятность этого события (т.е. $P(f_{выб})$) очень мала, то, в соответствии с приведенными рассуждениями, мы полагаем, что неправильно ее определили. Встает вопрос о том, что привело нас к ошибке. Вспоминаем, что мы находили вероятность в предположении справедливости проверяемой гипотезы. Именно это предположение и заставило нас считать вероятность встреченного значения очень малой. Поскольку опыт дает основания полагать, что в действительности вероятность не столь мала, остается отвергнуть нашу H_0.

Если же вероятность $P(f_{выб})$ достаточно велика для того, чтобы значение $f_{выб}$ могло встретиться практически, то мы полагаем, что у нас нет оснований сомневаться в справедливости проверяемой гипотезы. Мы принимаем последнюю, считаем, что она справедлива для генеральной совокупности.

Таким образом, право именоваться критерием функция f обретает в силу того, что именно величина ее значения играет определяющую роль в выборе одной из двух альтернатив: принятия гипотезы H_0 или отвержения ее.

Остался нерешенным вопрос о том, где граница между «малой» и «достаточно большой» вероятностью? Эта граница должна быть равна такому значению вероятности, относительно которого мы могли бы считать, что событие с такой (или с меньшей) вероятностью практически не может случиться – «не может быть, потому, что не может быть никогда». Это значение называют уровнем значимости принятия (отвержения) нуль-гипотезы и обозначают буквой α. Обычно полагают, что $\alpha = 0{,}05$, либо

$\alpha = 0{,}01$. Математическая статистика не дает нам правил определения α. Установить уровень значимости может помочь только практика.Конечно, этот уровень должен обусловливаться реальной задачей, тем, насколько социально значимым может явиться принятие ложной или отвержение истинной гипотезы (процесс проверки статистических гипотез всегда сопряжен с тем, что мы рискуем совершить одну из упомянутых ошибок). Если большие затраты (материальные, либо духовные) связаны с отвержением гипотезы, то мы будем стремиться сделать α как можно меньше, чтобы была как можно меньше вероятность отвержения правильной нуль-гипотезы. Если же затраты сопряжены с принятием гипотезы, то имеет смысл α увеличить, чтобы уменьшить вероятность принятия ложной гипотезы.

Теперь рассмотрим конкретную интересующую нас нулевую гипотезу: гипотезу об отсутствии связи между двумя изучаемыми номинальными переменными. Функция, выступающая в качестве описанного выше статистического критерия носит название «Хи-квадрат», обозначается иногда как X^2 (X – большое греческое «Хи»; подчеркнем, что далее будет фигурировать малая буква с тем же названием; и надо различать понятия, стоящие за этими обозначениями, что не всегда делается в ориентированной на социолога литературе).Определяется этот критерий следующим образом:

$$X^2 = \sum_{i,\,j} \left[\frac{(n_{ij}^{теор} - n_{ij}^{эмп})^2}{n_{ij}^{теор}} \right]$$

где $n_{ij}^{эмп}$ – наблюдаемая нами частота, стоящая на пересечении i-й строки и j-го столбца таблицы сопряженности (так называемая эмпирическая частота), а $n_{ij}^{теор}$ – та частота, которая стояла бы в той же клетке, если бы наши переменные были статистически независимы (т.е. та, которая отвечает пропорциональности столбцов (строк) таблицы сопряженности; она обычно называется теоретической, поскольку может быть найдена из теоретических соображений; иногда ее называют также ожидаемой часто-

той, поскольку действительно ее появление и ожидается при независимости переменных). Теоретическая частота обычно находится по формуле:

$$n_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n} . \quad (1)$$

Приведем доказательство этой формулы. Сделаем это не для приобщения читателя к математике, а для демонстрации того, как необходимо воспринимать частоты при грамотном анализе таблицы сопряженности. Доказательство, о котором мы говорим, является очень простым, и использующиеся в процессе его проведения принципы входят в число тех знаний, которыми должен владеть каждый социолог, анализирующий эмпирические данные.

Итак, мы утверждаем, что теоретическая частота отвечает той ситуации, когда являются независимыми два события – то, что первый признак принимает значение i, и то, что второй признак принимает значение j. Независимость же двух событий означает, что вероятность их совместного осуществления равна произведению вероятностей осуществления каждого в отдельности. Вычислим соответствующие вероятности для интересующего нас случая. Представляется очевидным, что эти вероятности хорошо оцениваются (имеются в виду выборочные оценки вероятностей с помощью относительных частот) следующим образом:

$$P(X = i. Y = j) = \frac{n_{ij}}{n}; \; P(X = i) = \frac{n_{i\cdot}}{n}; \; P(Y = j) = \frac{n_{\cdot j}}{n} .$$

Независимость наших событий означает справедливость соотношения:

$$P(X = i. Y = j) = P(X = i) \text{ w } P(Y = j)$$

или, учитывая введенные выше соотношения:

$$\frac{n_{ij}}{n} = \left(\frac{n_{i\cdot}}{n}\right) \times \left(\frac{n_{\cdot j}}{n}\right),$$

что легко преобразуется в доказываемое соотношение (1). Перейдем к описанию того, как «работает» наш критерий «Хи-квадрат».

Представим себе, что мы организуем бесконечное количество выборок и для каждой из них вычисляем величину X^2. Образуется последовательность таких величин:

$$X^2_{выб1}, \; X^2_{выб2}, \; X^2_{выб3}, \; ...$$

Очевидно, имеет смысл говорить об их распределении, т.е. об указании вероятности встречаемости каждого значения. В математической статистике доказано следующее положение: если наши признаки в генеральной совокупности независимы, то вычисленные для выборок значения X^2 приблизительно имеют хорошо изученное распределение, «имя» которого – $ч^2$ («Хи-квадрат», здесь используется малое греческое «хи»). Приблизительность можно игнорировать (т.е. считать, что величины X^2 распределены в точности по закону $ч^2$), если клетки тех выборочных частотных таблиц, на базе которых рассчитываются величины X^2, достаточно наполнены – обычно считают, что в каждой клетке должно быть по крайней мере 5 наблюдений. Будем считать, что это условие соблюдено.

Чтобы логика проверки нашей нуль-гипотезы стала более ясной, отметим, что при отсутствии связи в генеральной совокупности среди выборочных X^2, конечно, будут преобладать значения, близкие к нулю, поскольку отсутствие связи означает равенство эмпирических и теоретических частот и, следовательно, равенство X^2 нулю. Большие значения X^2 будут встречаться сравнительно редко – именно они будут маловероятны. Поэтому можно сказать, что большое значение X^2 должно приводить нас к утверждению о наличии связи, малое – об ее отсутствии.

Изученность распределения какой-либо случайной величины означает, что у нас имеется способ определения вероятности попадания каждого ее значения в любой заданный интервал – с помощью использования специальных вероятностных таблиц. Такие таблицы имеются и для распределения $ч^2$. Правда, надо помнить, что такое распределение не одно. Имеется целое семейство подобных распределений. Вид каждого зависит от размеров используемых частотных таблиц. Точнее, этот вид определяется так называемым числом степеней свободы df (degree

freedom) распределения:

$df = (r - 1)\ w\ (c - 1)$.

Итак, если в генеральной совокупности признаки независимы, то, вычислив число степеней свободы для интересующей нас матрицы, мы можем найти по соответствующей таблице вероятность попадания произвольного значения X^2 в любой заданный интервал. Теперь вспомним, что такое значение у нас одно – вычисленное для нашей единственной выборки. Обозначим его через $X^2_{выб}$. Описанная выше логика проверки статистической гипотезы превращается в следующее рассуждение.

Вычислим число степеней свободы df и зададимся некоторым уровнем значимости α. Найдем по таблице распределения $ч^2$ такое значение $X^2_{табл}$, называемое критическим значением критерия (иногда используется обозначение $X^2_{крит}$), для которого выполняется неравенство:

$$P(\xi \geq i\ X^2_{табл}) = \alpha$$

(ξ – обозначение случайной величины, имеющей распределение $ч^2$ с рассматриваемым числом степеней свободы).

Если $X^2_{выб} < X^2_{табл}$ (т.е. вероятность появления $X^2_{выб}$ достаточно велика), то полагаем, что наши выборочные наблюдения не дают оснований сомневаться в том, что в генеральной совокупности признаки действительно независимы – ведь, «ткнув» в одну выборку, мы встретили значение X^2, которое действительно вполне могло встретиться при независимости. В таком случае мы полагаем, что у нас нет оснований отвергать нашу нуль-гипотезу, и мы ее принимаем – считаем, что признаки независимы. Если же $X^2_{выб} \geq X^2_{табл}$ (т.е. вероятность появления $X^2_{выб}$ очень мала, т.е. меньше α), то мы вправе засомневаться в нашем предположении о независимости – ведь мы «наткнулись» на такое событие, которое вроде бы не должно было встретиться при этом предположении. В таком случае мы отвергаем нашу нуль-гипотезу – полагаем, что признаки зависимы.

Итак, рассматриваемый критерий не гарантирует наличие связи, не измеряет ее величину. Он либо говорит о том, что эмпирия не дает оснований сомневаться в отсутствии связи, либо, напротив, дает повод для сомнений.

2.3.1.3. Нормировка значений функции «Хи-квадрат»

Сами значения рассматриваемого критерия непригодны для оценки связи между признаками, поскольку они зависят от объема выборки и других обстоятельств, носящих, вообще говоря, случайный характер по отношению к силе измеряемой связи (о некоторых обстоятельствах подобного рода пойдет речь ниже). Так, величина критерия, например равная 30, может говорить о большой вероятности наличия связи, если в клетках исходной частотной таблицы стоят величины порядка 10, 20, 30, и о ничтожной вероятности того же, если рассматриваемые частоты равны 1000, 2000, 3000 и т.д. В таких случаях возникает необходимость определенной нормировки найденного значения критерия – такого его преобразования, которое устранит описанную зависимость от случайных (для оценки связи) факторов.

Подчеркнем, что здесь речь идет о принципиальном моменте, часто возникающем при использовании в социологии разного рода статистических критериев, индексов и т.д. Всегда необходимо выяснять, не отражает ли используемый показатель что-либо случайное по отношению к изучаемому явлению и в случае наличия такого отражения осуществлять соответствующую нормировку показателя.

Принято нормировку, подобную описанной, осуществлять таким образом, чтобы нормированные коэффициенты изменялись либо от –1 до +1 (если имеет смысл противопоставление положительной и отрицательной направленности изучаемого с помощью рассматриваемого индекса явления, в нашем случае – связи), либо от 0 до 1 (если выделение положительной и отрицательной направленности явления содержательно бессмысленно).

Подчеркнем, что приведение всех коэффициентов к одному и тому же интервалу является необходимым, но недостаточным условием, обеспечивающим возможность их сравнения. Если такого приведения не будет сделано, сравнение заведомо невоз-

можно. Но и при его осуществлении сравнение тоже может оказаться бессмысленным. Об этом пойдет речь в п.2.3.5.

Имеются разные подходы к требующейся нормировке. Наиболее известными являются такие, которые превращают критерий «Хи-квадрат» в известные коэффициенты, называемые обычно по именам впервые предложивших их авторов – Пирсона, Чупрова, Крамера. За этими коэффициентами утвердились постоянные обозначения, отвечающие первым буквам названных фамилий (коэффициент Чупрова отвечает немецкому tsch, коэффициент Крамера имеет два обозначения из-за известного различия букв, обозначающих звук «к» в разных языках):

$$P = \sqrt{\frac{X^2}{X^2 + n}}$$

$$T = \sqrt{\frac{X^2}{n\sqrt{(c-1)(r-1)}}}$$

$$K \text{ (или } C) = \sqrt{\frac{X^2}{n \times \min(c-1, r-1)}}$$

Опишем некоторые свойства этих коэффициентов. Начнем с тех, которые обычно оговариваются в литературе.

Все коэффициенты изменяются от 0 до 1 и равны нулю в случае полной независимости признаков (в описанном выше смысле). Как и критерий «Хи-квадрат», эти показатели являются симметричными относительно наших признаков: с их помощью нельзя выделить зависимую и независимую переменные, на основе их анализа нельзя говорить о том, какая переменная на какую «влияет».

Обычно в качестве недостатка коэффициента Пирсона P (предложенного в литературе первым) упоминается зависимость его максимальной величины от размера таблицы (максимум P достигается при $c = r$, но величина максимального значения изменяется с изменением числа категорий: при $c = 3$ значение P не

может быть больше 0,8, при $c = 5$ максимальное значение P равно 0,89 и т.д. [Интерпретация и анализ..., 1987, с.31]). Естественно, это приводит к возникновению трудностей при сравнении таблиц разного размера.

Отметим следующий немаловажный факт, очень редко рассматривающийся в ориентированной на социолога литературе.

Многие свойства рассматриваемых коэффициентов доказываются лишь при условии выполнения одного не всегда приемлемого для социологии предположения, состоящего в том, что за каждым нашим номинальным признаком «стоит» некая латентная (скрытая) непрерывная количественная (числовая) переменная.

Сделаем небольшое отступление по поводу используемых терминов. Все три определения к термину «переменная» требуют пояснения. Термин «латентная» употребляется здесь несколько условно. Обычно (в теории социологического измерения, например, в факторном, латентно-структурном анализе, многомерном шкалировании) под латентной переменной понимают признак, значения которого вообще не поддаются непосредственному измерению (например, путем прямого обращения к респонденту). Значения же нашей переменной мы измеряем самым непосредственным образом. Но получаем при этом номинальную шкалу, хотя и предполагаем, что между отвечающими этим значениям свойствами реальных объектов существуют отношения, достаточно сложные для того, чтобы можно было говорить об использовании интервальной шкалы (о соотношении между «богатством» реальных отношений, между эмпирическими объектами и типом шкал, использующихся при шкалировании этих объектов [Клигер и др., 1978; Толстова, 1998].

Термин «непрерывная» здесь употребляется в том смысле, что в качестве значения этой переменной может выступать любое рациональное число.

«Количественной» мы, в соответствии с традицией, называем переменную, значения которой получены по шкале, тип которой не ниже типа интервальной шкалы (о нашем отношении к подобному использованию терминов «качественный–количествен-

ный» уже шла речь в п.4.3 части I). Можно показать, что для таких шкал любое рациональное число может в принципе оказаться шкальным значением какого-либо объекта. Поэтому термины «количественный» и «непрерывный» часто употребляются как синонимы.)

Итак, мы полагаем, что задавая респонденту интересующий нас вопрос в анкете, мы как бы принуждаем его разбить весь диапазон изменения рассматриваемой количественной переменной на интервалы (количество которых равно числу значений нашей номинальной переменной) и указать, в каком из этих интервалов, по его мнению, находится оцениваемый объект. Внутри каждого интервала значения переменной становятся неразличимыми, между интервалами же определены лишь отношения совпадения – несовпадения (основное свойство номинальной шкалы). Когда исследователь имеет дело с двумя переменными такого рода (например, когда нас интересуют парные связи), то обычно предполагается еще и нормальность соответствующего двумерного распределения.

Именно такого предположения придерживался Пирсон, когда в начале века вводил свой коэффициент. Он доказал, что P равно тому предельному значению обычного коэффициента корреляции между латентными переменными, к которому этот коэффициент стремится при безграничном увеличении количества градаций рассматриваемых признаков. Очевидно, что без указанного предположения было бы совершенно неясно, как подобное свойство коэффициента P можно проинтерпретировать.

Для исправления указанного выше недостатка коэффициента Пирсона (зависимости его максимально возможного значения от размеров таблицы сопряженности) Чупров ввел коэффициент T, названный его именем. Но и T достигает единицы лишь при $c = r$, и не достигает при $c \neq r$. Может достигать единицы независимо от вида таблицы коэффициент Крамера K. Для квадратных таблиц коэффициенты Крамера и Чупрова совпадают, в остальных случаях $K > T$.

Мы перечислили те свойства рассматриваемых коэффициентов, которые часто упоминаются в литературе. Из редко упоминающихся свойств можно назвать еще один свойственный всем коэффициентам недостаток – зависимость их величины от соотношений маргинальных частот анализируемой таблицы сопряженности (подчеркнем очень важный момент – вычисляя теоретические частоты при расчете критерия «Хи-квадрат», мы пользуемся маргинальными суммами, полагая, что имеем дело с их «генеральными» значениями, что, вообще говоря, не всегда отвечает реальности).

О том, как измерять связь между номинальными признаками с помощью критерия «Хи-квадрат», можно прочесть в работах [Елисеева, 1982; Елисеева, Рукавишников, 1977, с.82-89; Интерпретация и анализ..., 1987, с.31-32; Лакутин, Толстова, 1990; Паниотто, Максименко, 1982, с.65-84; Рабочая книга социолога, 1983, с.169-172, 190 (с учетом того, что на с.169 речь идет о таких теоретических частотах, которые являются частотами таблицы сопряженности, отвечающей случаю статистической независимости рассматриваемых номинальных переменных); Статистические методы..., 1979, с.117-120; Толстова, 1990а, с.54-57]

Перейдем к описанию таких коэффициентов парной связи, которые основаны на других априорных моделях, на другом понимании сути этой связи.

2.3.2. Коэффициенты связи, основанные на моделях прогноза

2.3.2.1. Выражение представлений о связи через прогноз

Включение понятия прогноза в представление о связи между номинальными признаками представляется разумным: наверное, трудно возражать против того, чтобы признаки считались связанными, если знание значения одного признака позволяет улучшить прогноз значения другого. Поясним это на гипотетическом примере, который ниже мы будем неоднократно «эксплуатировать». Заодно уточним только что сформулированное суждение.

Предположим, что мы изучаем жителей некоторого крупного города N от 20 лет и старше и что нас интересует связь между признаком «возраст» , рассматриваемым нами как номинальный, и дихотомическим признаком со значениями «студент» – «не студент».

(Напомним два принципиальных для социологии момента. Во-первых, определение типа шкалы для таких, казалось бы, «понятных» признаков, как возраст, далеко не всегда является ясным делом; причиной тому служит то, что их значения, как правило, интересуют исследователя не сами по себе, а лишь как показатели некоторых латентных переменных. Во-вторых, здесь мы отвлекаемся от сложной проблемы разбиения диапазона изменения непрерывного признака – предполагаем, что это сделано каким-либо адекватным решаемой задаче образом.)

Предположим, что распределение изучаемой совокупности по возрасту приблизительно равномерно, например, такое, какое изображено на рисунке 14.

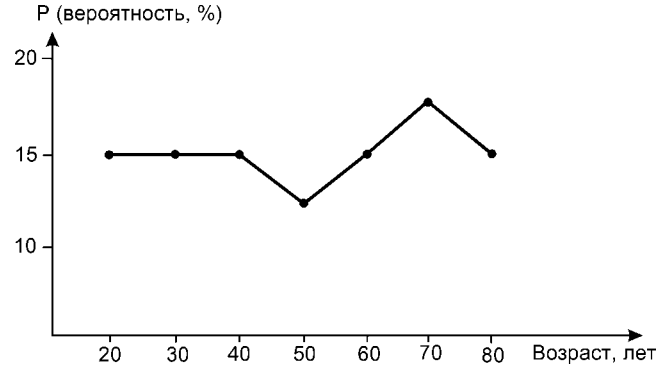

Рис.14. Гипотетическое распределение по возрасту жителей города N старше 20 лет

Интуитивно ясно, что в такой ситуации мы вряд ли сможем хорошо прогнозировать возраст респондента. Выбрав наугад (случайным образом) произвольного человека, мы примерно с

одинаковой степенью уверенности можем полагать, что он имеет любой возраст: вероятность «наткнуться» на 20-летнего юношу такая же, как и на 80-летнего старика (подчеркнем своеобразие понимания нами термина «прогноз» – речь идет просто о том, что мы можем сказать о значении возраста для случайно выбранного респондента).

Другое дело, если мы рассмотрим только студентов. Ясно, что их распределение по возрасту будет резко отличаться от общего. Например, будет иметь вид, изображенный на рисунке 15.

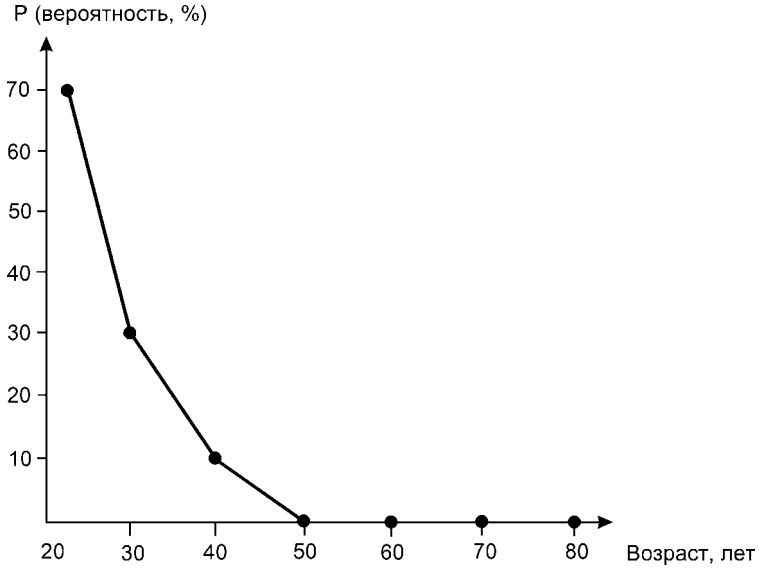

Рис.15. Гипотетическое распределение по возрасту студентов города N старше 20 лет

Ясно, что теперь, случайным образом отобрав человека (студента), мы с уверенностью 90% (90 = 70 + 20) будем полагать, что его возраст не превысит 30 лет, вероятность же «попасть» на человека старше 40 лет практически равна нулю.

Итак, фиксировав значение «студент» второго рассматриваемого нами признака, мы явно улучшили возможность прогноза

возраста жителей города. Наверное, на основе этого было бы разумно сделать вывод о наличии связи между признаком «возраст» и признаком «быть студентом». Подчеркнем, что для того, чтобы сделать этот вывод, мы *сравнили* безусловное распределение признака «возраст» (см. рис.14) с его условным распределением (см. рис.15), когда условие состоит в фиксации значения «студент» второго признака. Возможность хорошего прогноза на основе знания условного распределения сама по себе (без ее сравнения с возможностью прогноза по безусловному распределению) ни о какой связи еще не говорит. Так, изучая только студентов, мы не можем говорить о связи пола и возраста на основе того, что, отобрав только девушек, мы можем хорошо прогнозировать их возраст. Ведь всего вероятнее, столь же хороший прогноз может быть осуществлен и для юношей, и для студентов вообще (т.е. для безусловного распределения). О соотношении безусловного и условного распределений при изучении связей см. также [Лакутин, Толстова, 1990].

Итак, будем считать, что смысл рассматриваемых (прогнозных) коэффициентов на интуитивном уровне ясен. Все такие коэффициенты должны служить мерой улучшения качества прогноза значения одного признака за счет получения сведений о значении другого признака по сравнению с тем случаем, когда последнее значение неизвестно. Такие коэффициенты и будем называть опирающимися на модель прогноза.

Для того, чтобы можно было практически пользоваться высказанными предположениями, необходимо их формализовать. Другими словами, необходимо четко понять, что такое прогноз и как именно на основе частотной таблицы мы можем судить о различии возможности прогноза для соответствующих условных и безусловных распределений. Формализация может быть разной. И, в первую очередь, неоднозначно может пониматься сам термин «прогноз». Те известные коэффициенты связи, которые мы намереваемся рассмотреть, отличаются друг от друга как раз способом формализации этого понятия. Но прежде, чем переходить к описанию некоторых прогнозных коэффициентов, напомним, что проблема формализации содержательных

представлений о «прогнозной» связи, вообще говоря, не исчерпывается рассуждениями о понимании прогноза и оценке его качества. Отметим также следующие три немаловажные момента.

Во-первых, глобальные коэффициенты связи по существу являются «усреднениями» всевозможных локальных коэффициентов. И подобные «усреднения» могут пониматься по-разному, выражаться разными формулами. Это также обусловливает наличие разных коэффициентов связи.

Во-вторых, возможность осуществления прогноза значений одного признака по значениям другого существенно зависит от того, значения какого признака прогнозируются. Скажем, значения первого могут хорошо прогнозироваться по значениям второго, а значения второго по значениям первого – очень плохо. Приведем простой, несколько утрированный пример. Пусть частотное распределение значений двух признаков имеет вид, представленный в таблице 12.

Таблица 12

Таблица сопряженности, иллюстрирующая несимметричность понятия «прогноз»

X	Y		
	1	2	3
1	0	0	10
2	0	0	10
3	0	20	0

Ясно, что по значению X мы легко предсказываем значение Y. Обратное же не имеет места: если признак Y равен 3, то X с одинаковым успехом (с равной вероятностью) может принимать значения 1 или 2. В таком случае возникает вопрос о построении коэффициентов, несимметричных относительно рассматриваемых признаков или, как говорят, коэффициентов, отражающих направленную связь – скажем, говорящих о том, появляется ли у нас новая информация о втором признаке при фиксации значения первого, но ничего не говорящих об обратной зависимости.

Актуальной является задача усреднения таких направленных коэффициентов для оценки ненаправленной связи. Обоснование

соответствующей необходимости – примерно такое же, как обоснование необходимости использования глобальных коэффициентов наряду с локальными: с одной стороны, не имея коэффициентов направленной связи, мы можем упустить, не заметить важные причинно-следственные отношения, но, с другой, – когда направленные связи не очень значимы, мы можем «за деревьями» не увидеть леса» – не уловить того, что, хотя каждая направленная связь не очень велика, в целом нельзя игнорировать взаимодействие рассматриваемых признаков.

О терминах: когда говорят о прогнозе значения признака Y по признаку X, то X называют независимой переменной, а Y – зависимой.

Перейдем к описанию наиболее известных коэффициентов, основанных на моделях прогноза.

2.3.2.2. Коэффициенты, основанные на модальном прогнозе

Формализуем понятие прогноза следующим образом. Выбирая произвольный объект и зная распределение рассматриваемого признака (условное или безусловное), считаем, что для выбранного объекта этот признак принимает то значение, которое имеет максимальную вероятность, встречается с максимальной частотой (т.е. модальное значение). Такой прогноз называется модальным. Чтобы стал ясен содержательный смысл рассматриваемого прогноза, приведем формулы соответствующих коэффициентов. Но сначала отметим, что таких коэффициентов три: два отражают возможные направленные связи, а третий является их усреднением. Эти коэффициенты обычно обозначаются буквами л с индексами: $л_r$ – отражающий «влияние» строкового признака на столбцовый; $л_c$ – отражающий «влияние» столбцового признака на строковый, л – усредненный коэффициент.

Рассмотрим формулу для $л_r$, (для $л_c$ рассуждения совершенно аналогичны). Будем использовать те же обозначения, которые были задействованы выше.

$$\lambda_r = \frac{\sum_{i=1}^{r} \max_j n_{ij} - \max_j n_{.j}}{n - \max_j n_{.j}} \qquad (2)$$

Выражение $\max_j n_{ij}$ означает наибольшую частоту в i-й строке.

Выражение $\max_i n_{ij}$ – наибольшую столбцовую маргинальную частоту.

Поясним смысл формулы (2) на примере. Пусть частотная таблица имеет вид:

Таблица 13

Пример частотной таблицы, использованной для расчета коэффициента $Л_r$

X	Y			Итого
	1	2	3	
1	0	20	30	50
2	5	15	30	50
3	40	5	5	50
Итого	45	40	65	150

Наибольшая частота в первой строке матрицы равна 30, во второй – тоже 30, в третьей – 40. Максимальный маргинал по столбцам – 65. Общее количество объектов в выборке – 150. Значит, имеет место равенство:

$$\lambda_r = \frac{(30 + 30 + 40) - 65}{150 - 65} = 0{,}41$$

Рассмотрим безусловное распределение признака Y. Отвечающие ему частоты – это маргиналы по столбцам рассматриваемой матрицы: 45, 40, 65. Модальная частота – 65. Значит, выбрав случайным образом какой-либо объект, мы, прогнозируя

для него значение Y, в соответствии с нашими представлениями о прогнозе, должны сказать, что упомянутое значение равно 3 (именно это значение является модой). Ясно, что, поступая так и перебирая последовательно всех респондентов, мы дадим правильный прогноз в 65 случаях и ошибемся в (150–65) случаях (заметим, что доля (вероятность) ошибки будет равна $\frac{150 - 65}{150}$). Именно эта разность стоит в знаменателе нашей формулы.

Итак, для безусловного распределения качество нашего прогноза можно оценить с помощью величины (150–65). Улучшится ли прогноз при переходе к условным распределениям того же признака? Попытаемся ответить на этот вопрос.

Пусть X = 1. Соответствующее условное распределение Y определяется частотами первой строки нашей матрицы: числами 0, 20, 30. Значит, перебирая 50 респондентов с первым значением X и делая для каждого прогноз в соответствии с нашими правилами, мы не ошибемся в 30 случаях. При X = 2 количество верных предположений тоже будет равно 30. При X =3 – 40. Общее количество правильных прогнозов во всех условных распределениях будет равно (30+30+40). По сравнению с «безусловным» случаем оно возрастет на ((30+30+40) – 65) единиц. А это – числитель выражения для $л_г$.

Итак, в числителе формулы (2) отражена величина того суммарного прироста количества правильных прогнозов, который возникает за счет перехода от перебора объектов, «сваленных в одну кучу» («куча» отвечает безусловному распределению), к перебору последовательно по «слоям» (отвечающим условным распределениям). Эта величина отражает суть коэффициента. Знаменатель же формулы (2) использован для нормировки (знаменатель равен значению числителя, получающемуся, когда суммарный прогноз по условным распределениям будет стопроцентным). Потребность в таковой возникает в силу тех же причин, которые были обсуждены нами при рассмотрении критерия «Хи-квадрат»: без нормировки величина коэффициента будет зависеть от размера выборки, значений конкретных частот и т.д.

Теперь, чтобы закончить вопрос о том, как в рассматриваемом случае формализуются естественные представления о связи, необходимо затронуть проблему «усреднения» всевозможных связей типа «альтернатива–альтернатива». Способ усреднения очевиден. Он как бы двуступенчат. Рассматривая какое-либо из наших условных распределений, мы говорим о прогнозе, учитывая сразу все возможные значения Y, не анализируя отдельно, насколько зафиксированное значение X может быть связано с тем или иным значением Y (в п.2.3.2.3 мы увидим, как такая связь может быть прослежена).

Переходя к общей формуле, мы суммируем показатели качества прогноза для всех условных распределений, игнорируя то, что для одного значения X этот прогноз может быть хорошим, а для другого – плохим.

В заключение обсуждения вопроса о $л_г$ опишем некоторые его свойства.

Имеют место неравенства: $0 \le л_г \le 1$. Коэффициент приближается к 1 по мере того, как в каждой строке объекты все более концентрируются в одной клетке, т.е. прогноз значения Y для условных распределений становится все лучше. Нетрудно проверить, что $л_г = 1$, если

$$\sum_{i=1}^{r} \max_j n_{ij} = n$$

,

и что это, в свою очередь, может быть верным лишь в случае, когда в каждой строке частотной таблицы существует только одна отличная от нуля частота, т.е. когда по значению признака X мы можем однозначно судить о значении признака Y (но не обратно!).Чем ближе значение $л_г$ к 1, тем лучше такое предсказание и сильнее связь (в рассматриваемом понимании) между переменными.

$л_г = 0$, если максимальные частоты в строках приходятся на один и тот же столбец. Это имеет место даже в том случае, если все остальные элементы частотной таблицы близки к нулю, т.е. если фактически имеется «хорошая» связь (а отнюдь не отсутствие связи,

как это должно было бы быть для нулевого значения хорошего коэффициента связи). И это является существенным недостатком рассматриваемого коэффициента.

Как мы уже отмечали, все приведенные рассуждения справедливы и для коэффициента л, служащего показателем связи, если зависимая и независимая переменные меняются местами, и вычисляющегося по формуле:

$$\lambda_c = \frac{\sum_{j=1}^{c} \max_i n_{ij} - \max_i n_{i.}}{n - \max_i n_{i.}}$$

Для измерения по тому же принципу ненаправленной связи показатели рассматриваемых направленных связей усредняются. Это делается разными способами. Самый простой:

$$\lambda = \frac{\lambda_r + \lambda_c}{2}$$

Итак, подведем итог обсуждению рассмотренных коэффициентов. Правила их построения определяют отвечающее модальному классу значение зависимого признака (Y) как оценку этого значения для произвольно взятого объекта. Если оценка делается без знания значения независимого признака (X) , то значением, предсказываемым для всех объектов, является модальное значение безусловного распределения зависимого. Если же оценка делается на основе знания значения X, то прогноз осуществляется отдельно для объектов, обладающих этим значением, на основе выявления моды соответствующего условного распределения Y. Величина $л_r$ ($л_c$) говорит об уменьшении (за счет осуществления перехода от безусловного распределения к набору условных) ошибки осуществленного с единичной вероятностью предсказания о том, что объект обладает модальным значением Y.

Приведем несколько утрированный пример. Рассмотрим, как может измеряться связь между национальностью (X) и цветом волос (Y). Предположим, что Вы являетесь продавцом косметики и Вам для того, чтобы подготовиться к общению с покупателем, желательно заранее знать цвет его волос. Представим себе, что Вы арендовали помещение в вузе и к Вам в комнату по очереди (в случайном порядке) входят за покупкой студенты. Допустим также, что Вы знаете безусловное распределение всех студентов рассматриваемого вуза по цвету волос, и в соответствии с этим распределением количество блондинов, брюнетов и шатенов примерно одинаково, но шатенов несколько больше, чем остальных. Вы пользуетесь правилом: перед входом покупателя выставляете товар, рассчитанный на модальное значение признака «цвет волос» (в нашем случае – на шатенов).

Теперь представим себе две ситуации.

В первой Вы ничего не знаете о национальности входящего к Вам студента. Наверное,в таком случае, приготовив товар для шатенов, Вы в почти двух третях возможных случаев совершите ошибку: к Вам с одинаковой вероятностью в любой момент может войти и блондин, и брюнет, и шатен. Торговля заведомо будет неэффективной.

А во второй ситуации Вы сумели организовать дело так, что сначала к Вам по очереди (снова в случайном порядке) входят учащиеся в вузе китайцы, затем – финны, потом – русские. Очевидно, эффективность Вашей торговли возрастет: зная, что сегодня к Вам придут китайцы, Вы готовите товар, рассчитанный только на брюнетов, если придут финны – на блондинов, если русские – на шатенов. Конечно, Вы и тут будете ошибаться, но уже в гораздо меньшей степени, чем раньше. Другими словами, Ваш прогноз улучшится. А это и означает наличие связи между национальностью и цветом волос. Чем в большей мере прогноз улучшился, тем сильнее связь.

Описанный прогноз называют модальным или оптимальным. Коэффициенты чаще всего называют коэффициентами Гуттмана [Интерпретация и анализ..., 1987; Статистические методы..., 1979], Гудмена [Паниотто, Максименко, 1982] или л-коэффициентами [Рабочая книга..., 1983].

2.3.2.3. Общее представление о пропорциональном прогнозе

Представленное понимание прогноза не является единственно возможным. Более того, его нельзя признать наилучшим. Прогноз здесь очень груб, приблизителен. Используя достижения теории вероятностей, к определению понятия прогноза можно подойти более тонко. Опишем еще один подход. На нем тоже базируется целый ряд известных коэффициентов связи (например, коэффициент Валлиса [Интерпретация и анализ..., 1987; Статистические методы..., 1979]). Принцип их «действия» по существу является тем же, что и принцип л-коэффициентов. Отличие состоит только в понимании процедуры прогноза. Мы не будем эти коэффициенты описывать, поскольку такое описание требует использования довольно сложных формул, но ничего не дает принципиально нового для понимания отражаемой с помощью этих коэффициентов связи.

Итак, что же такое пропорциональный прогноз? Опишем его суть с помощью примера.

Предположим, что мы имеем дело с частотной таблицей 13. Рассмотрим безусловное распределение Y. Обратимся к схематичному изображению ситуации в терминах столь часто фигурирующих в литературе по теории вероятностей урн и заполняющих их шаров. Возьмем 150 шаров, на 45 из них напишем цифру 1, на 40 – цифру 2, на 65 – цифру 3 и погрузим все шары в урну, перемешав их. Правило прогноза выглядит очень просто: берем случайного респондента, опускаем руку в урну и вытаскиваем тот шар, который случайно же нам попался. То, что на нем написано, и будет прогнозным значением признака Y для выбранного респондента. Аналогичным образом поступаем и для каждого условного распределения. Конечно, реализовать такой подход можно и без шаров с урнами, но суть должна сохраниться: то, что чаще встречается в исходной совокупности, должно чаще попадаться в наши руки при вытаскивании шаров. К примеру, в соответствии с первым условным распределением (X = 1, первая строка частотной таблицы), у нас отсутствуют респонденты, для которых Y = 1. Не будут попадаться нам и шары с единицей, поскольку количество таких шаров равно 0. В соответствии с третьим распределением (X = 3) значения 2 и 3 признака Y встречаются одинаково часто и в 8 раз реже значения 1. И вероятность встречаемости шаров с цифрами 2 и 3 будет одинаковой и в 8 раз меньше вероятности встречаемости шара с 1.

Описанный прогноз называется пропорциональным. Хотя соответствующее правило, на первый взгляд, довольно сложно, оно позволяет предсказывать значение зависимого признака с большей надежностью, чем правило модального прогноза. Это часто используется в самых разных прогнозных алгоритмах.

2.3.3. Коэффициенты связи, основанные на понятии энтропии

Семейство коэффициентов, к рассмотрению которых мы переходим, основаны на такой модели связи, которая очень близка по своему содержательному смыслу к прогнозным моделям. В основе этих коэффициентов также лежит сравнение безусловного распределения с условными (условие – фиксация значения независимого признака X). Но сравнение это ведется не с точки зрения того, насколько при переходе от безусловного распределения к условным меняется качество возможного прогноза, а с точки зрения изучения изменения степени неопределенности рассматриваемых распределений. Здесь мы, как и в п.1.3.5, вступаем в область теории информации и будем использовать ее терминологию.

2.3.3.1. Условная и многомерная энтропия

Вернемся к рассмотренному нами в п.1.3.5 понятию энтропии.

По аналогии с энтропией распределения одного признака, определяется энтропия двумерного распределения:

$$H(X, Y) = -\sum_{i,j} P(X = i\,.\,Y = j) \times \log P(X = i\,.\,Y = j).$$

Точка внутри скобок означает конъюнкцию соответствующих событий, одновременное их выполнение. Если ввести обозначения, аналогичные использованным выше: $P_{ij} = P(X=i.\ Y=j)$, то же соотношение запишется в виде:

$$H(X, Y) = -\sum_{i,j} P_{ij} \times \log P_{ij}.$$

Точно так же можно определить энтропию любого многомерного распределения.

Необходимо дать определение еще одного очень важного для нас понятия – так называемой условной энтропии:

$$H(Y/X) = -\sum_{i} P_i \cdot H(Y/X=i) =$$

$$= \sum_{i} P_i \sum_{j} P(Y=j/X=i) \times \log P(Y=j/X=i). \quad (3)$$

Можно доказать следующие свойства энтропии.

$H(X, X) = H(X)$; $H(X,Y) = H(X) + H(Y/X)$;

$H(X,Y) \leq H(X) + H(Y)$;

равенство в последнем соотношении появляется только тогда, когда X и Y статистически независимы, т.е. когда выполняется уже обсужденное нами соотношение: $P_{ij} = P_i w P_{j\cdot}$.

В определенном смысле противоположным понятию энтропии является понятие информации, к рассмотрению которого мы переходим. (Отметим, что говоря об информации в сочетании с энтропией, мы вступаем в сферу мощного научного направления – теории информации. Решающим этапом в становлении этой теории явилась публикация ряда работ К.Шеннона в 1948–1949 годах.)

Приобретение информации сопровождается уменьшением неопределенности, поэтому количество информации можно измерять количеством исчезнувшей неопределенности, т.е. степенью уменьшения энтропии. Ниже речь пойдет об информации, содержащейся в одном признаке (случайной величине) относительно другого признака. Поясним смысл этого понятия более

подробно, по существу используя другой язык для описания того же, о чем шла речь выше [Яглом, Яглом, 1960, с.78].

Вернемся к величине H(Y), характеризующей степень неопределенности распределения Y или, говоря несколько иначе, степень неопределенности опыта, состоящего в том, что мы случайным образом отбираем некоторый объект и измеряем для него величину Y.

Если $H(Y) = 0$, то исход опыта заранее известен. Большее или меньшее значение H(Y) означает большую или меньшую проблематичность результата опыта. Измерение признака X, предшествующее нашему опыту по измерению Y, может уменьшить количество возможных исходов опыта и тем самым уменьшить степень его неопределенности. Для того, чтобы результат измерения X мог сказаться на опыте, состоящем в измерении Y, необходимо, чтобы упомянутый результат не был известен заранее. Значит, измерение X можно рассматривать как некий вспомогательный опыт, также имеющий несколько возможных исходов. Тот факт, что измерение X уменьшает степень неопределенности Y, находит свое отражение в том, что условная энтропия опыта, состоящего в измерении Y, при условии измерения X оказывается меньше (точнее, не больше) первоначальной энтропии того же опыта. При этом, если измерение Y не зависит от измерения X, то сведения об X не уменьшают энтропию Y, т.е. $H(Y/X) = H(Y)$. Если же результат измерения X полностью определяет последующее измерение Y, то энтропия Y уменьшается до нуля:

$$H(Y/X) = 0.$$

Таким образом, разность

$$I(X, Y) = H(Y) - H(Y/X) \quad (4)$$

указывает, насколько осуществление опыта по измерению X уменьшает неопределенность Y, т.е. сколько нового мы узнаем об Y, произведя измерение X. Эту разность называют *количеством информации* относительно Y, содержащейся в X (в научный обиход термин был введен Шенноном).

Приведенные рассуждения о смысле понятия информации очевидным образом отвечают описанной выше логике сравнения безусловного и условных распределений Y. В основе всех информационных мер связи (а о них пойдет речь ниже) лежит та разность, которая стоит в правой части равенства (4). Но именно эта разность и говорит о различии упомянутых распределений. Нетрудно понять и то, каким образом здесь происходит усреднение рассматриваемых характеристик всех условных распределений (напомним, что в качестве характеристики распределения у нас выступает его неопределенность, энтропия). По самому своему определению (см. соотношение (3)) выражение H (Y/X) есть взвешенная сумма всех условных энтропий (каждому значению признака X отвечает своя условная энтропия Y:

,

причем каждое слагаемое берется с весом, равным вероятности появления соответствующего условного распределения, т.е. вероятности P_i. Другими словами, можно сделать вывод, что для выборки величина H (Y/X) – это обычное среднее взвешенное значение условных энтропий.

О возможных способах нормировки разности (H(Y) – – H (Y/X)) пойдет речь далее, поскольку рассматриваемые ниже коэффициенты именно этой нормировкой фактически и отличаются друг от друга.

В заключение настоящего параграфа опишем некоторые свойства информации.

I (X,Y) – функция, симметричная относительно аргументов, поскольку, как нетрудно показать, имеет место соотношение:

$$I(X,Y) = H(X) + H(Y) - H(X, Y),$$

а функция H (X,Y) симметрична по самому своему определению. Другими словами, количество информации, содержащейся в X относительно Y, равно количеству информации в Y относительно X, т.е. соотношение (4) эквивалентно соотношению

$$I(X,Y) = H(X) - H(Y/X).$$

Перейдем к описанию мер связи, основанных на понятии энтропии.

2.3.3.2. Смысл энтропийных коэффициентов связи. Их формальное выражение

Поскольку понятие энтропии является как бы обратной стороной понятия информации, то энтропийные коэффициенты в литературе нередко называют информационными. Мы эти два термина будем использовать как синонимы.

Переходя к обсуждению конкретных информационных мер связи, прежде всего отметим, что в качестве такой меры может служить I(X, Y). Как мы уже отметили, это – симметричная (значит, – ненаправленная) мера. Из приведенных выше свойств энтропии следуют следующие свойства названной меры:

$$I(X,Y) \geq 0$$

$$\sum_j P(Y = j/X = i) \leq \log P(Y = j/X = i)$$

где равенство достигается тогда и только тогда, когда X и Y статистически независимы и

$$I(X, X) = H(X).$$

Широко известны и направленные меры связи:

$$C_{X/Y} = \frac{I(X,Y)}{H(X)} \quad \text{и} \quad C_{Y/X} = \frac{I(Y,X)}{H(Y)}.$$

Первый из этих коэффициентов можно интерпретировать как относительное приращение информации об X, возникающее за счет знания Y [Миркин, 1980. С. 103]. Относительность возникает в результате соотнесения такого приращения с первоначальной неопределенностью распределения X. Аналогично интерпретируется и второй коэффициент.

Коэффициенты C называют асимметричными коэффициентами неопределенности, коэффициентами нормированной информации [Елисеева, Рукавишников, 1977, с.91]. Нетрудно проверить справедливость следующих соотношений [Елисеева, Рукавишников, 1977; Статистические методы..., 1979]:

$0 \le C_{X/Y} \le 1;$

$C_{X/Y} = 0$ если и только если переменные X и Y независимы; $C_{X/Y}=1$, если и только если X однозначно определяется значением Y (т.е. если можно говорить о детерминистской зависимости X от Y; о том, что мера разнообразия X определяется мерой разнообразия Y единственным образом, т.е. о полной связи).

Ясно, что аналогичными свойствами обладает и коэффициент $C_{Y/X}$.

Соответствующий симметризованный коэффициент нормированной информации вводится следующим образом [Елисеева, Рукавишников, 1977, с.95]:

$$R(Y,X) = \frac{I(X,Y)}{0{,}5\,(H(X)+H(Y))}.$$

Часто используется также коэффициент Райского:

$$R(Y,X) = \frac{I(X,Y)}{H(X,Y)}.$$

Нетрудно проверить, что он обладает свойствами, аналогичными сформулированным выше свойствам коэффициентов C: заключен в интервале от 0 до 1, в 0 обращается тогда и только тогда, когда признаки статистически независимы, а в 1 – тогда и только тогда, когда признаки полностью детерминируют друг друга.

Введенные информационные меры связи во многом похожи на обычный коэффициент корреляции. Но они имеют одно преимущество перед последним: из того, что коэффициент корреляции равен 0, вообще говоря, не следует статистическая независимость рассматриваемых признаков, а из равенства 0 рассмотренных информационных мер связи – следует.

Описание информационных мер связи можно найти в [Миркин, 1980; Статистические методы..., 1979; Елисеева, Рукавишников, 1977].

2.3.4. Коэффициенты связи для четырехклеточных таблиц сопряженности. Отношения преобладаний

Четырехклеточные таблицы – это частотные таблицы, построенные для двух дихотомических признаков. Встает вопрос – надо ли изучать эти таблицы отдельно? Ведь они представляют собой частный случай всех возможных таблиц сопряженности. Выше мы обсуждали коэффициенты, которые можно использовать для анализа любой частотной таблицы, в том числе и для четырехклеточной. Однако ответ на наш вопрос положителен. Причин тому несколько.

Во-первых, многие известные коэффициенты для четырехклеточных таблиц оказываются равными друг другу. И по крайней мере надо знать об этом, чтобы не осуществлять заведомо ненужные выкладки.

Во-вторых, оказыватся, что именно в анализе четырехклеточных таблиц можно увидеть нечто полезное для социолога, но не высвечивающееся на таблицах большей размерности.

В-третьих, с помощью анализа специальным образом организованных четырехклеточных таблиц оказывается возможным перейти от изучения глобальных связей к изучению локальных и промежуточных между первыми и вторыми (о промежуточных связях мы говорили в п.2.2.1).

Итак, рассмотрим два дихотомических признака – X и Y, принимающие значения 0 и 1 каждый, и отвечающую им четырехклеточную таблицу сопряженности (табл.14).

Ниже будем использовать пример, когда рассматриваются два дихотомических признака – пол (1 – мужчина, 0 – женщина) и курение (1 – курит, 0 – не курит) (табл.15).

Таблица 14

Общий вид четырехклеточной таблицы сопряженности

X	Y		Итого
	1	0	
1	a	b	a + b
0	c	d	c + d
Итого	a + c	b + d	a + b + c + d

Буквы в клетках обозначают соответствующие частоты

Таблица 15

Пример четырехклеточной таблицы сопряженности

Курение	Пол		Итого
	м	ж	
Курит	80	4	84
Не курит	10	6	16
Итого	90	10	100

Данные таблицы 15 говорят о том, что в нашей совокупности имеется 90 мужчин, из которых 80 человек курят, и 10 женщин, среди которых 4 человека курящих и т.д.

Все известные коэффициенты связи для четырехклеточных таблиц основаны на сравнении произведений ad и bc. Если эти произведения близки друг к другу, то полагаем, что связи нет. Если они совсем не похожи – связь есть. Основано такое соображение на том, что равенство $ad = bc$ эквивалентно равенству $\frac{a}{c} = \frac{b}{d}$, что, в свою очередь, означает пропорциональность столбцов (строк) нашей частотной таблицы, т.е отсутствие статистической связи. Чем более отличны друг от друга указанные произведения, тем менее пропорциональны столбцы (строки) и, стало быть, тем больше оснований имеется у нас полагать, что переменные связаны. Для обоснование этого утверждения могут быть использованы те же рассуждения, что были приведены выше. А именно, можно показать, что разница между наблюдаемой и теоретической частотой для левой верхней клетки нашей четырехклеточной частотной таблицы (нетрудно проверить, что наличие или отсутствие связи для такой таблицы определяется содержанием единственной клетки – при заданных маргиналах частоты, стоящие в других клетках, можно определить однозначно) равна величине [Кендалл, Стьюарт, 1973, с.722]:

$$D = \frac{ad - bc}{n}$$

Коэффициенты, основанные на описанной логике, могут строиться по-разному. Но всегда они базируются либо на оценке разности $(ad - bc)$, либо на оценке отношения $\frac{ad}{bc}$. В первом случае об отсутствии связи будет говорить близость разности к нулю, во втором – близость отношения к единице. Естественно, ни разность, ни отношение не могут служить искомыми коэффициентами в «чистом» виде, поскольку их значения зависят от величин используемых частот. Требуется определенная нормировка. И, как мы уже оговаривали выше, желательно, чтобы искомые показатели связи находились либо в интервале от –1 до 1, либо – от 0 до 1. Возможны разные ее варианты. Это обуславливает наличие разных коэффициентов – показателей связи для четырехклеточных таблиц. Рассмотрим два наиболее популярных коэффициента.

Коэффициент ассоциации Юла:

$$Q = \frac{ad - bc}{ad + bc}$$

и *коэффициент контингенции*

$$\Phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}.$$

Коротко рассмотрим их основные свойства.

Оба коэффициента изменяются в интервале от –1 до +1 (значит, для них имеет смысл направленность связи; о том, что это такое в данном случае, пойдет речь ниже). Обращаются в нуль в случае отсутствия статистической зависимости, о котором мы говорили выше (независимость признаков связана с пропорциональностью столбцов таблицы сопряженности). А вот в единицу (или –1) эти коэффициенты обращаются в разных ситуациях. Они схематично отражены на рисунке 16.

Свойства коэффициентов:	Q = 1	Q = –1	Ф = 1	Ф = –1
Отвечающие им виды таблиц	a 0 c d a b 0 d	0 b c d a b c 0	a 0 0 d	0 b c 0
	(а)	(б)	(в)	(г)

Рис.16. Схематическое изображение свойств коэффициентов Q и Ф

Таким образом, мы видим, что Q обращается в 1, если хотя бы один элемент главной диагонали частотной таблицы равен 0. Для обращения же в 1 коэффициента Ф необходимо обращение в 0 обоих элементов главной диагонали. Нужны ли социологу оба коэффициента? Покажем, что каждый из них позволяет выделять свои закономерности. Или, как мы говорили выше, – за каждым из них стоит своя модель изучаемого явления, свое понимание связи, выделение как бы одной стороны того, что происходит в реальности. Постараемся убедить читателя, что социолога должны интересовать обе эти стороны.

Предположим, что в нашем распоряжении имеется лишь коэффициент Ф, и мы даем задание ЭВМ для каких-то массивов данных выдать нам все такие четырехклеточные таблицы, для которых этот коэффициент близок к единице (может быть, мы хотим найти все те признаки, для которых имеется связь для респондентов некоторой фиксированной совокупности, а, может быть – изучаем, для каких совокупностей респондентов имеется сильная связь между какими-то конкретными признаками). ЭВМ выдаст нам набор таблиц типа (в) или (г). Мы будем знать, к

примеру, что имеются группы респондентов, для которых имеется сильная связь между полом и курением: все мужчины курят, а все женщины не курят (что довольно распространено) или наоборот – все женщины курят, а мужчины – нет (что имеет место, скажем, для некоторых индейских племен). Но мы «не заметим», что для каких-то групп все мужчины курят, в то время как среди женщин встречаются и курящие, и не курящие, либо все женщины не курят, хотя мужчины ведут себя по-разному – могут и курить, и не курить (случай (а)). Думается, что не требует особого доказательства утверждение о том, что социолог, не умеющий выискивать подобные ситуации, рискует много потерять. Аналогичное утверждение справедливо и относительно ситуаций, обозначенных буквой (б).

Другими словами, не используя коэффициент Q, социолог рискует не заметить интересующие его закономерности. Перефразируя сказанное выше и вспомнив, что связь также имеет отношение и к прогнозу, отметим, что эти не замеченные закономерности отвечают ситуациям, когда мы по одному значению первого признака можем прогнозировать значение второго, а по другому значению не можем: скажем, зная, что респондент – мужчина, мы с полной уверенностью можем сказать, что он курит, а зная, что респондент – женщина – никакого прогноза, вообще говоря, делать не можем (нижняя таблица случая (а)). Вряд ли можно сомневаться, что выявление и такой «половинчатой» возможности прогноза для социолога может быть полезной.

Рассмотрим теперь вопрос: не можем ли мы обойтись без коэффициента Ф? Представляется очевидным отрицательный ответ на него: выявляя значимые ситуации только с помощью Q, мы можем «за деревьями не увидеть леса» – не заметить, что в отдельных случаях мы может прогнозировать не только по одному значению того или иного признака, но и по другому тоже.

Описанное различие между коэффициентами Q и Ф нашло свое отражение в терминологии. Та связь, которую отражает Q, была названа *полной,* а та, которую отражает Ф, – *абсолютной.*

Еще раз определим эти виды связи, несколько видоизменив формулировку. Для этого вспомним, что, зная маргиналы

четырехклеточной таблицы сопряженности, о связи между двумя дихотомическими признаками можно судить по одной частоте. Чаще всего для этого используют n_{11}. Обозначим отвечающие этой частоте значения наших признаков через A и B. Например, A означает «мужчина», а B – «курит». В таком случае говорят, что связь между A и B полная, если все A являются одновременно B, несмотря на то, что не все B являются одновременно A (все мужчины курят, но не все курящие являются мужчинами). Если же все A являются одновременно B и все B являются одновременно A (т.е. если все мужчины курят и все курящие – мужчины), то связь называется абсолютной.

Иногда для обозначения тех же свойств рассматриваемой связи используют иную терминологию – говорят, что Q измеряет одностороннюю связь, а Ф – двустороннюю.

Поясним теперь, в чем смысл знака рассматриваемой связи. Для этого заметим, что приведенные выше рассуждения можно переформулировать, говоря не о том, что все A являются одновременно B, а о том, что свойства A и B сопрягаются друг с другом (таблица *сопряженности* потому так и названа, что ее придумали для того, чтобы изучать, какие значения разных признаков «ходят» вместе, сопрягаются друг с другом). Термины «положительный» и «отрицательный», используемые для характеристики связи, носят весьма относительный характер: «положительность» означает, что какое-то значение первого признака сопрягается с одним значением другого, а «отрицательность» – с другим (при наличии положительной связи все мужчины курят, и при наличии отрицательной – все мужчины не курят).

Однако сказанное становится весьма нечетким утверждением при отсутствии нулевых клеток в таблице сопряженности. Например, трудно понять, с каким значением признака «курит – не курит» сопрягается мужской пол, если данные представлены таблицей 16.

С одной стороны, среди курящих больше женщин, чем мужчин. И среди женщин больше курящих, чем некурящих. Но правильно ли будет сказать, что свойство «курит» сопрягается с женским полом? Ведь если среди мужчин в 2,5 раза (50:20) больше курящих,

Таблица 16

Частотная таблица для демонстрации отношения преобладаний

Курение	Пол		Итого
	М	Ж	
Курит	50	90	140
Не курит	20	40	60
Итого	70	130	200

чем некурящих, то среди женщин – лишь в 2,25 раза (90:40). Строгое определение положительной и отрицательной связи можно дать с помощью введения понятия *отношения преобладаний* [Rudas,1998]:

$$\lambda = \frac{50:20}{90:40}$$

или, в общем случае (обозначения – как в таблице 14):

$$\lambda = \frac{a:c}{b:d}.$$

Если отношение преобладания больше единицы, то связь называется *положительной,* если меньше единицы – то *отрицательной.* (Отношение преобладания обобщается на многомерный случай, о чем коротко пойдет речь в п.2.3.5).

И еще об одном очень важном моменте необходимо сказать. Если мы, используя обозначения 0 и 1 для значений наших признаков, будем интерпретировать эти обозначения как настоящие числа, то, как нетрудно проверить, вычисленный по обычным правилам коэффициент корреляции между признаками окажется равным Ф. Будучи обобщенным, этот факт имеет огромное значение для анализа данных. Дело в том, что одним из популярных способов создания возможности использования числовых математико-статистических методов для анализа номинальных (нечисловых!) данных является так называемая дихотомизация последних: замена (по определенным правилам) одного номинального признак таким количеством дихотомических, принимающих значения 0 и 1, сколько в этом признаке

альтернатив, и дальнейшая «работа» с этими 0 и 1 как с обычными числами. Этот подход не имеет строгого математического обоснования. Его «оправдание» состоит в том, что все числовые статистики, рассчитанные по обычным правилам, оказывается возможным разумно проинтерпретировать. Именно пример этого мы и видели выше: коэффициент корреляции, вычисленный для 0 и 1, оказался разумной величиной, совпал с Ф. Вернемся к этому в п.2.6.3.

О коэффициентах связи для четырехклеточных таблиц можно прочесть в [Интерпретация и анализ..., 1987, с.29-30; Лакутин, Толстова, 1990, 1992; Паниотто, Максименко, 1982, с.84-93; Рабочая книга..., 1983, с.189; Статистические методы..., 1979, с.116-117; Libetrau, 1989].

2.3.5. Проблема сравнения коэффициентов связи

Заканчивая обсуждение вопроса о коэффициентах связи типа «признак–признак», необходимо упомянуть актуальную для социологии проблему сравнения всех этих коэффициентов. Однако здесь мы не будем ее подробно обсуждать, отослав читателя к соответствующей литературе [Елисеева, Рукавишщников, 1982, с.89-101; Интерпретация и анализ..., 1987, с.34-36; Лакутин, Толстова, 1990, 1992; Миркин, 1980. с.94-109; Паниотто, Максименко, 1982, с.124-125; Рабочая книга..., 1983, с.191-192].

Отметим лишь очень коротко несколько отдельных моментов.

Любой критерий сравнения, как всякий подход к математическому анализу данных, основан на предположениях о том, что реальности адекватны некоторые формальные построения, отражающие определенные аспекты интерпретации исходных данных. Другими словами, для того, чтобы можно было говорить о сравнении, необходимо заранее сформировать некоторую модель того, что мы понимаем под схожими (несхожими) коэффициентами.

Наиболее обоснованное теоретически и часто использующееся в статистической литературе основание для сравнения

рассматриваемых коэффициентов базируется на обсужденном выше предположении о том, что за каждым номинальным признаком стоит некоторая латентная непрерывная количественная переменная. Коротко говоря, суть соответствующих подходов заключается в следующем. Исследователь моделирует с помощью ЭВМ некоторую «генеральную совокупность», описываемую двумя непрерывными переменными с заданным коэффициентом корреляции между ними. Затем упомянутые переменные искусственным образом превращаются в номинальные, из «генеральной» совокупности формируется множество выборок и для каждой из них подсчитываются подлежащие сравнению коэффициенты. Когда выборок организуется достаточно много, появляется возможность сравнения «поведения» отдельных коэффициентов друг с другом.

Сказанное в предыдущих параграфах свидетельствует о том, что все рассмотренные коэффициенты различны. За каждым стоит своя модель, свое понимание этой связи. Вопрос о том, какова же истинная связь между переменными, если такой-то коэффициент равен 0,7, а такой-то – 0,2, не имеет смысла. В описанной ситуации можно сказать только то, что связь в первом смысле (смысле, отвечающем первому коэффициенту) более высока, чем связь во втором смысле. И для того, чтобы найти «истинную» связь, надо использовать целый набор коэффициентов. Каждый их них как бы отвечает отдельной стороне «истины». А для того, чтобы «истина», как бриллиант, засверкала всеми своими гранями, необходимо иметь эти грани перед глазами все сразу, «поворачивая» нашу связь в разные стороны.

Однако имеет смысл сказать не только о различии, но и о сходстве разных коэффициентов. Если посмотреть на них с другой стороны, окажется, что не так уж сильно они расходятся друг с другом. И это не случайно – все-таки речь идет о разных способах формализации одного и того же явления – интуитивно понимаемой связи между переменными. Действительно, можно показать (и это в определенной мере демонстрировалось выше), что так или иначе, в разной степени, но все коэффициенты основаны на представлении о том, что существование связи между двумя

признаками означает одновременное соблюдение следующих условий: сильное отклонение от пропорциональности столбцов (строк) исходной таблицы сопряженности; улучшение качества прогноза значений одного признака при получении информации о значении другого; тот факт, что определенные значения одного признака «любят» встречаться вместе с определенными значениями другого признака. Однако относительно последнего обстоятельства можно заметить следующее (приведем цитату из [Кендалл, Стьюарт, 1973, с.724]):

«Следует обратить внимание на то, что статистическая связь отличается от связи в обычном смысле. В повседневной речи мы говорим, что A и B связаны, если они достаточно часто встречаются вместе, а в статистике они считаются связанными только в том случае, если A встречается относительно чаще среди B, чем среди не-B. Если 90% курящих страдают плохим пищеварением, то мы не можем сказать, что курение и плохое пищеварение связаны, пока не будет показано, что среди некурящих страдают плохим пищеварением менее, чем 90%». Последнее обстоятельство связано с тем, о чем пойдет речь в следующем параграфе.

2.3.6. Учет фактической многомерности реальных связей. Многомерные отношения преобладаний

Коснемся очень важной для практики проблемы, связанной со сравнением коэффициентов не друг с другом, а с некоторыми другими подходами к измерению связи между переменными.

Актуальность многомерных связей в социологии

В реальности двумерных связей практически не существует. Все связи многомерны. Приведем определения.

Связь между тремя переменными называется *трехмерной*, если характер связи между любыми двумя из них зависит от того, каково при этом значение третьей переменной. Связь между четырьмя переменными называется *четырехмерной,* если ее характер для

любых трех признаков зависит от того, каково при этом значение четвертой переменной и т.д. Надеемся ясно, как определяется понятие связи любой размерности.

Многомерность реальных зависимостей заставляет относиться с большой осторожностью к значениями рассмотренных выше парных коэффициентов связи. На это обстоятельство обращают внимание многие исследователи. Поясним это.

В работе [Миркин, 1985, с.18-20] приводится пример того, как фиксация значения третьей переменной обусловливает «возникновение» связи между двумя переменными. Опишем его.

Изучалась зависимость между наличием в семьях пылесоса и холодильника. Исходная частотная таблица имела вид:

	П	¬П	
X	560	840	1400
¬X	240	360	600
	800	1200	2000

Зависимость явно отсутствует, поскольку столбцы (строки) таблицы пропорциональны: $\frac{560}{240} = \frac{840}{360} = \frac{1400}{600} = \frac{7}{3}$. Таблицу пересчитали отдельно для двух выделенных среди изучаемой совокупности респондентов групп – для семей с высоким (Д) и низким (¬Д) уровнем дохода. Получились следующие две частотные таблицы:

Для Д:

	П	¬П	
X	520	300	820
¬X	80	100	180
	600	400	1000

Для ¬Д:

	П	¬П	
X	40	540	580
¬X	160	260	420
	200	800	1000

В обоих случаях связь присутствует (пропорциональности строк здесь явно нет). Более того, для первой таблицы она положительна (значение «X» сопрягается со значением «П»: семьи,

имеющие холодильник, как правило, имеют и пылесос), а для второй – отрицательна (значение «Х» сопрягается со значением ¬П: семьи, имеющие холодильник, чаще всего не могут купить пылесос).

Вспомнив определение положительной и отрицательной связи через отношение преобладания (п.2.3.4), то же самое выразим более строго. В таблице, отвечающей высокому доходу Д отношение преобладания $\frac{520:80}{300:100} = \frac{13}{6}$ больше единицы, а в таблице, отвечающей низкому доходу аналогичное отношение $\frac{40:160}{540:260} = \frac{13}{108}$ – меньше единицы.

Аналогичный пример, когда статистическая независимость между двумя признаками превращается в зависимость при фиксации значения третьего признака приводится в работе [ДА-система..., 1997, с.181-182].

В [Типология и классификация..., 1982] приводится заимствованный у Лазарсфельда пример того, как фиксация значения третьего признака, напротив, приводит к исчезновению первоначальной двумерной связи.

Речь идет о связи между чтением двух журналов А и Б. Исходная частотная таблица имеет вид (А – респондент читает журнал А, ¬А – не читает, то же для журнала Б):

	А	¬А	
Б	260	240	500
¬Б	140	360	500
	400	600	1000

Столбцы не пропорциональны: $\frac{260}{140} = \frac{13}{7} \neq \frac{240}{360} = \frac{2}{3}$.

Далее вводится новая переменная – образование респондента (В – высокое, ¬В – низкое). Соответствующие таблицы выглядят так:

Для В:

	А	¬А	
Б	240	160	400
¬Б	60	40	100
	300	200	500

Для ¬В:

	А	¬А	
Б	20	80	100
¬Б	80	320	400
	100	400	500

Нетрудно проверить, что столбцы обеих таблицы пропорциональны, т.е. зависимость в обоих случаях отсутствует. Связь исчезла. В таких случаях говорят, что уровень образования является переменной, объясняющей связь между чтением двух рассматриваемых журналов (здесь мы имеем дело с основным положением, лежащим в основе процесса измерения латентных переменных – с лазарсфельдовской аксиомой локальной независимости; эта аксиома лежит в основе латентно-структурного анализа).

В работе [Аптон, 1982] рассматриваемая проблема обсуждается в исторической ретроспективе. В частности, приводится пример так называемого парадокса Симпсона (1951 год). Приведем соответствующие данные. Исходная таблица имела вид:

	В	¬В	
А	495	805	1300
¬А	405	295	700
	900	1100	2000

В ней наблюдается явная отрицательная связь: отношение преобладаний $\frac{495:405}{805:295} = 0{,}45$ – меньше единицы (значение А имеет бьльшую тенденцию встречаться с ¬В, чем с В). А в тех двух таблицах, которые получаются в результате фиксирования значения третьего дихотомического признака С оба отношения преобладаний больше единицы, т.е. говорят о положительной связи. Эти таблицы выглядят так:

Для C:

	B	¬B	
A	95	800	895
¬A	5	100	105
	100	900	1000

Для ¬C:

	B	¬B	
A	400	5	405
¬A	400	195	595
	800	200	1000

Соответствующие же отношения преобладаний равны:

$$\frac{95:5}{800:100} = \frac{19}{8} \quad \text{и} \quad \frac{400:400}{5:195} = 39{,}0.$$

Многомерные отношения преобладаний

Как это уже неоднократно упоминалось, все приведенные соотношения в реальности теряют смысл из-за того, что мы имеем дело лишь со статистическими закономерностями. Что значат выражения типа: «при фиксации значения третьей переменной связь между первыми двумя исчезла»? Ведь и при наличии связи отклонение от пропорциональности столбцов носит лишь относительный характер, и при отсутствии связи у нас все же, как правило, пропорциональность не «чистая». Чтобы справиться с неопределенностью, можно использовать отношения преобладаний, введенные нами в п.2.3.4. Однако требуется их обобщить на многомерный случай. Сделаем это.

Вообще говоря, отношения преобладаний могут быть определены для таблиц любой размерности, в том числе и для одномерных, т.е. для линейных частотных распределений (правда, мы предполагаем, что имеем дело с дихотомическими признаками). Чтобы ввести строгое определение отношения преобладаний, введем новые обозначения.

Сначала предположим, что в нашем распоряжении имеется лишь один признак. Тогда будем обозначать через P_1 долю объектов, обладающих первым его значением, а через P_2 – вторым. Соответствующее отношение преобладания первого порядка, выражаемое формулой

$$\lambda_1 = \frac{P_1}{P_2},$$

естественно, будет обозначать, во сколько раз объем первого множества больше (меньше) второго. Если отношение преобладания больше 1, говорим о положительном преобладании, если меньше – об отрицательном.

Теперь будем считать, что у нас два дихотомических признака. Через P_{11} обозначим долю объектов с первым значением первого признака и первым значением второго, через P_{12} – с первым значением первого и вторым значением второго и т.д. Двумерная частотная таблица приобретет вид:

P_{11} P_{12}

P_{21} P_{22}

Легко видеть, что отношение преобладания второго порядка (определенное нами в п.2.3.4 и названное там просто отношением преобладания) конструируется следующим образом.

Фиксируем первое значение второго признака и рассчитываем для соответствующей частотной таблицы отношение преобладания первого порядка:

$$\frac{P_{11}}{P_{21}}$$

То же делаем при фиксации второго значения второго признака:

$$\frac{P_{12}}{P_{22}}.$$

Отношением преобладания второго порядка называется отношение первой дроби ко второй.

$$\lambda_2 = \frac{P_{11} : P_{21}}{P_{12} : P_{22}}.$$

2.6. Анализ связей типа «признак – группа признаков»: номинальный регрессионный анализ (NPA)

2.6.1. Общая постановка задачи

Вспомним некоторые рассуждения, использованные нами выше (п.2.2) в процессе осмысления предложенной классификации методов изучения связей между номинальными переменными. Мы подчеркивали, что в большинстве реальных задач исследователь не должен следовать ставшему традиционным ограничению круга используемых математических методов только известными коэффициентами парной связи. При этом описывалось две совокупности факторов, обусловливающих необходимость перехода к другим методам (рис.20).

Во-первых, имеет смысл «рассыпать» все рассматриваемые признаки на отдельные альтернативы и затем, «склеивая» их разными способами, искать такие сочетания значений исходных признаков, которые определяют те или иные связи, то или иное «поведение» респондентов (анализ фрагментов таблиц сопряженности, алгоритмы последовательных разбиений и т.д.).

Во-вторых, имеет смысл объединять отдельные признаки друг с другом, искать такие их сочетания, которые в каком-то смысле детерминируют другие признаки и их сочетания (как мы увидим ниже, в регрессионном анализе речь пойдет о детерминации среднего уровня этих «других» признаков). К соответствующим рассмотрениям мы и перейдем в настоящем параграфе. Проанализируем ту группу методов (или задач, мы говорили о том, что задачи для нас в определенном смысле отождествляются с методами), которая при класификации задач была символически обозначена нами как методы типа «признак – группа признаков». Сюда относится регрессионный анализ, к рассмотрению которого мы и переходим.

Сначала для простоты изложения рассмотрим случай, когда у нас имеется только два признака – X и Y – и нас интересует

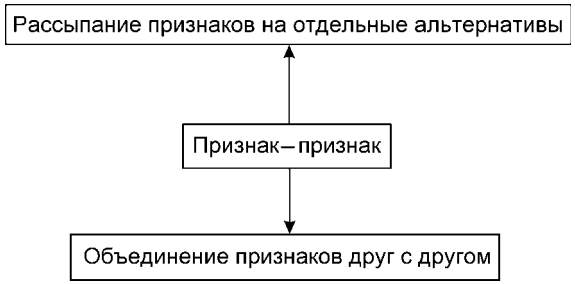

Рис.20. Схематичное выражение причин, обусловливающих необходимость перехода от традиционных коэффициентов парной связи к другим методам анализа связей

зависимость между ними. Другими словами, сначала предположим, что наша «группа признаков» состоит из одного признака – X (потом перейдем к случаю, когда вместо одного X фигурируют несколько признаков). Мы знаем, что о связи между признаками говорит соответствующий коэффициент корреляции: чем ближе значение модуля этого коэффициента к 1, тем более сильна эта связь, т.е. тем с большей уверенностью мы можем полагать, что с ростом значений одного признака растут (если коэффициент корреляции положителен) или убывают (если коэффициент корреляции отрицателен) значения другого (напомним, что коэффициент корреляции измеряет линейную связь между переменными; отметим, однако, что приводимые рассуждения справедливы и для других коэффициентов связи, например, для корреляционного отношения, дающего возможность оценить криволинейную связь). Но при этом мы совершенно не можем сказать о том, в какой степени возрастет значение Y, если значение X увеличится, скажем, на 1. А ситуации здесь могут быть весьма разными.

Приведем пример, рассмотрев зависимость между производственным стажем человека и его зарплатой. Предположим, что мы имеем дело с двумя крайними ситуациями, отраженными на рисунках 21*а* и 21*б*. В обоих случаях соответствующие коэффициенты корреляции близки к 1 (обе совокупности точек-

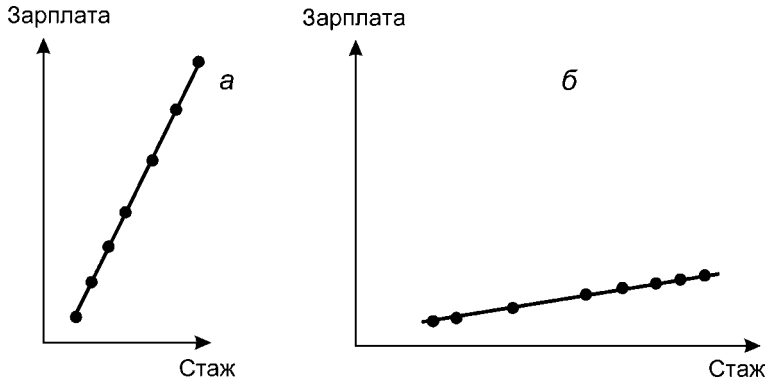

Рис.21. Примеры сильных линейных связей, определяющих разный прогноз

объектов лежат на прямых линиях, отвечающих нашей зависимости). На первом из них прямая идет резко вверх. Поэтому даже при небольшом увеличении X признак Y резко возрастет. В случае же наличия связи, изображенной на втором рисунке, прямая близка к горизонтали. Поэтому даже при значительном росте X значение Y почти не изменится. Другими словами, на основании наших двух картинок мы получим прогнозы совершенно различного характера. И совершенно ясно, что этого никак нельзя узнать лишь на основе вычисления соответствующих коэффициентов корреляции.

Итак, для того, чтобы делать прогноз о том, как изменится значение Y при том или ином изменении значения X, нам желательно знать, как говорят, форму связи между этими переменными, т. е. желательно найти функцию вида Y = f (X). Подчеркнем, что отношение между X и Y несимметрично: речь идет именно о зависимости второй переменной от первой, именно о возможности прогноза значения Y от X, а не наоборот.

В данном случае для обозначения X и Y используются те же термины, о которых шла речь в начале п.2.5.3.1. Однако для той ситуации, когда речь идет о нахождении формы зависимости Y от X, употребляется еще несколько пар терминов: независимые переменные называют *входными, экзогенными, внешними*, а зависимая – *выходной, эндогенной, внутренней*. Представляется важным

правильное понимание причин использования такой терминологии.

Поиск функции f предполагает разработку определенной модели связи между переменными, опирающуюся на априорные знания исследователя (так, ниже мы будем говорить в основном о линейной модели, о *линейном* регрессионном анализе). Найденная с помощью регрессионной техники зависимость – это тоже некоторая модель реальности – модель, в соответствии с которой и находятся значения Y на основе информации о значениях признака X.

Независимые признаки (X) потому и можно назвать независимыми, что они не зависят от этой модели. Эти признаки как бы поступают на ее «вход», являются внешними по отношению к ней, берутся «со стороны». Они определяют конкретный вид искомой зависимости, но не определяются ею. Прогнозируемые же значения зависимой переменной (Y) полностью определяются моделью (то, насколько они близки к реальности, зависит от качества модели), служат ее «выходом», являются ее порождением. Они внутренне по отношению к ней.

Особенно осторожно надо использовать словосочетания «признак–причина» и «признак–следствие», о чем мы уже говорили в п.2.1.3.

2.6.2. Повторение основных идей классического регрессионного анализа, рассчитанного на так называемые «количественные» признаки

Сначала для простоты и возможности геометрического изображения основных положений регрессионного анализа предположим, что у нас всего две переменные: X и Y (соответственно, независимая и зависимая). С помощью рассматриваемого подхода осуществляется поиск зависимости вида Y = f(X). Однако это выражение для результата регрессионного анализа носит условный характер: искомая зависимость не функциональна, а статистична, является закономерностью «в среднем», она «неточна». Поясним, в чем именно состоят такие усредненность и «неточность».

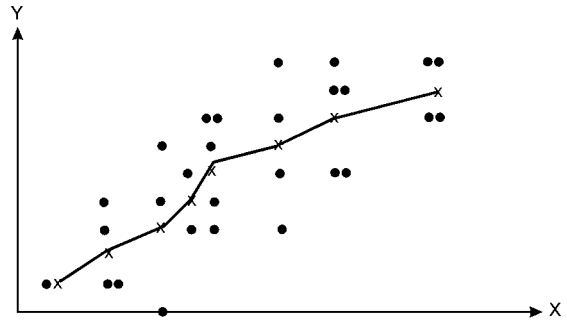

Рис.22. Принципиальная схема линии регрессии

В качестве независимой переменной фигурируют условные средние значения Y (каждое такое среднее вычисляется для конкретного значения независимой переменной X; соответствующая точка на графике обозначена крестиком)

Прежде всего обратим внимание читателя на то, что для социологических данных типична ситуация, когда одному значению X соответствует множество значений Y. Эта ситуация схематично изображена на рисунке 22 (пока обращаем внимание только на черные кружки).

Встает вопрос: какую именно зависимость мы хотим вычислить? Как искомая кривая (а мы хотим, чтобы каждому значению независимой переменной отвечало одно значение зависимой, т.е. чтобы искомой связи отвечала какая-то одномерная линия) должна «пробиваться» через изображенное на рисунке облако точек?

Ответ представляется естественным: подсчитаем для каждого значения X среднее арифметическое значение всех отвечающих ему значений Y и будем изучать зависимость от X именно таких средних. Соответствующие точки на нашем рисунке обозначены крестиками. Для них вид искомой зависимости четко «просматривается». Другими словами, интересующая нас статистическая зависимость будет иметь вид:

$$\overline{Y}_X = f(X). \qquad (8)$$

Вспомним, что на рисунке 22 отражена выборочная ситуация, в то время как в действительности нас интересует то, что делается в генеральной совокупности. Рассмотрение последней предполагает, что переменные непрерывны, имеют бесконечное число значений. Соотношение (8) для генеральной совокупности превращается в следующее:

$$\mu(Y/X) = f(X), \qquad (9)$$

(где м – знак математического ожидания – меры средней тенденции для генеральной совокупности; напомним, что среднее арифметическое, является лишь «хорошей» выборочной оценкой математического ожидания). Такая функция называется *функцией регрессии* Y по X (иногда говорят об *уравнении регрессии,* либо о *регрессионной зависимости*). Ее график называется *линией регрессии.* Подчеркнем, что соотношение (9) предполагает, что при каждом фиксированном значении X значения Y суть значения некоторой случайной величины. Это означает следующее.

Фиксируя какое-либо значение X, равное, например, X_i (т.е. рассматривая совокупность объектов, обладающих этим значением), мы имеем дело с некоторым условным распределением Y (которое образуют значения зависимой переменной Y, вычисленные для объектов, обладающих значением X_i признака X). Это распределение имеет свое математическое ожидание и дисперсию. Именно это математическое ожидание фигурирует в левой части равенства (9). Это математическое ожидание лежит на линии регрессии (рис.23).

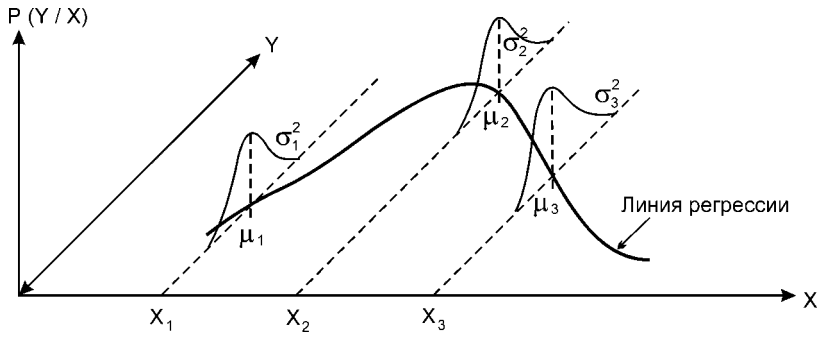

Рис.23. Статистические предположения, лежащие в основе регрессионного анализа

Условные распределения зависимой переменной Y нормальны. Их математические ожидания μ_1, μ_2, μ_3 лежат на линии регрессии; дисперсии σ_1^2, σ_2^2, σ_3^2 равны

μ_1, μ_2, μ_3 – математические ожидания тех условных распределений переменной Y, которые получаются при фиксации, соответственно, значений X_1, X_2, X_3 переменной X. Ясно, что с помощью линии регрессии хорошо можно осуществлять тот прогноз, который является основной целью поиска зависимости Y от X: эта линия говорит о том, насколько изменится среднее значение Y при том или ином изменении значения X. Будем говорить в таком случае об *изменении Y в среднем.*

Точность, с которой линия регрессии Y по X передает изменение Y в среднем при изменении X, измеряется дисперсией величины Y, вычисленной для каждого значения X:

Пусть σ_1^2, σ_2^2, σ_3^2 – значения дисперсий, вычисленных для условных распределений переменной Y, получающихся при фиксации, соответственно, значений X_1, X_2, X_3 переменной X.

Обычно предполагается, что описанные условные распределения зависимой переменной Y нормальны, а дисперсии этих распределений равны: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$. Именно такая ситуация отражена на рисунке 23. При равенстве дисперсий говорят, что условные распределения удовлетворяют свойству *гомоскедастичности*. Попытаемся коротко пояснить смысл этого свойства.

Ясно, что чем меньше условные дисперсии Y, т.е. чем меньше разброс зависимого признака в условных распределениях, тем больше можно верить прогнозу значений этого признака, осуществляемому с помощью уравнения регрессии. Напротив, большой разброс может полностью лишить нас возможности делать прогноз: утверждение о том, что для такого-то X_i переменная Y в среднем равна соответствующему условному среднему, не будет иметь никакой практической ценности из-за того, что бессмысленным станет сам расчет средней величины (в п.1.2 мы говорили о том, что для осмысленности средней требуется однородность изучаемой совокупности объектов, отсутствие большого разброса по рассматриваемому признаку). Можно говорить о качестве найденной регрессионной зависимости,

связывая его именно с описанной возможностью прогноза. Тогда при условных дисперсиях, равных одной и той же величине σ, это качество может быть строго определено: при большой σ оно будет плохим, при малой – хорошим. А если разбросы при разных X разные? Тогда для одних значений X уравнение регрессии будет хорошим, при других – плохим. Представляется, что при практическом использовании такого уравнения могут возникнуть неприятности. Отсюда – требование гомоскедастичности.

Теперь обсудим вопрос о том, как найти конкретный вид функции регрессии f. На помощь приходит то, что линия регрессии обладает замечательным свойством: среди всех действительных функций f минимум математического ожидания $\mu(Y - f(X))^2$ достигается для функции $f(X) = \mu(Y/X)$. Поясним смысл этого утверждения, обратившись к выборочной ситуации, представленной на рисунке 24.

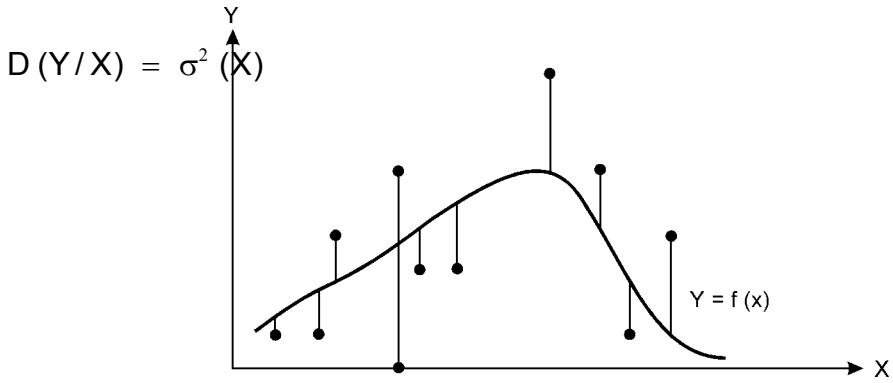

Рис.24. Отклонения ординат рассматриваемых точек от произвольной функции

Рассмотрим заданную совокупность точек – моделей изучаемых объектов и произвольную функцию f (X). Вертикальные отрезки – отклонения ординат рассматриваемых точек графика от этой функции. Средняя величина квадратов длин этих отрезков – это и есть выборочная оценка математического ожидания $\mu(Y - f(X))^2$.

Для того, чтобы лучше понять способ вычисления величин рассмотренных отрезков, покажем, в чем он состоит, на примере одной точки, имеющей произвольные координаты (X,Y) в нашем признаковом пространстве. Обратимся к рисунку 25.

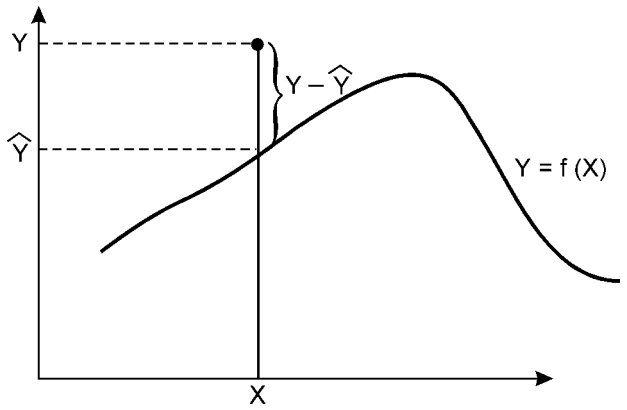

Рис.25. Способ определения отклонения точки (X, Y) от произвольной функции Y = f (X)

X координата рассматривамого объекта (на рисунке он обозначен точкой) по оси X; Y – его же координата по оси Y; $\widehat{Y}$ – ордината точки, принадлежащей графику функции Y = f (X) и имеющей по оси X ту же координату, что и наш объект.

Сумма $\Sigma (Y - \widehat{Y})^2$ (суммирование осуществляется по всем рассматриваемым объектам) и есть та величина, которую надо минимизировать для того, чтобы получить выборочное представление линии регрессии. Символически процесс такой минимизации можно выразить следующим образом:

$$\Sigma (Y - \widehat{Y})^2 \in \min \qquad (10)$$

$\widehat{Y}$ – это как бы теоретическое, модельное значение зависимой переменной. Это то значение, которое мы имели бы, если бы после всех расчетов пользовались найденной функцией Y = f (X) как основой для прогноза.

В соответствии с сформулированным выше свойством линии регрессии, можно сказать, что минимальной эта сумма будет в

том случае, если рассматриваемая функция Y = f (X) является выборочным представлением искомой линии регрессии. Другими словами, указанному выборочному представлению отвечает та функция f (X), для которой указанная выше сумма минимальна.

Итак, чтобы найти выборочную линию регрессии, необходимо как бы «перебрать» все возможные функции Y = f (X), для каждой вычислить указанную сумму квадратов и остановиться на той функции, для которой эта сумма минимальна.

Рассматриваемый способ поиска f (X), носит название *метода наименьших квадратов*. (Отметим, что этот метод очень часто используется при расчете самых разных статистических закономерностей. Так, он задействован в одном из известных методов шкалирования – методе парных сравнений [Толстова, 1998].)

Чтобы смысл метода наименьших квадратов стал яснее, заметим, что чем меньше величина указанной выше суммы квадратов, тем с большим основанием рассматриваемую функцию можно считать близкой одновременно ко всем рассматриваемым точкам. Эта функция в каком-то смысле служит моделью всего «облака» точек. Это можно проиллюстрировать с помощью рисунка 26.

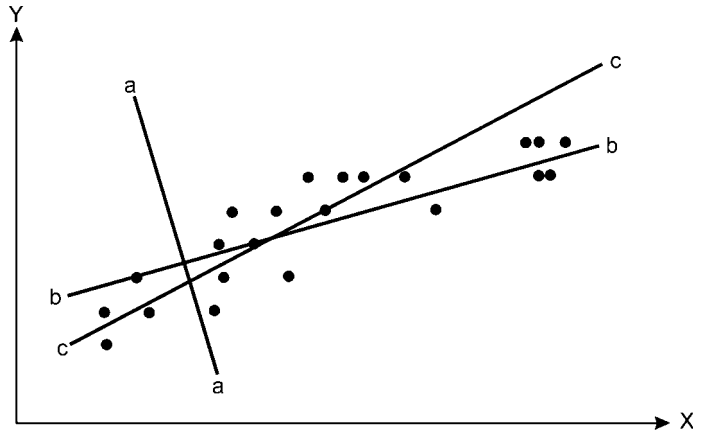

Рис.26. Иллюстрация проблемы выбора прямой линии, наилучшим образом отвечающей линии регрессии

Ясно, что прямая «aa» заведомо не может минимизировать рассматриваемую сумму: она совсем не отражает наше облако точек. А вот относительно прямых «bb» и «cc» вряд ли «на глаз» можно определить, какая из них лучше. Чтобы ответить на этот вопрос, необходимо использовать метод наименьших квадратов.

Очевидно, перебрать все мыслимые функции невозможно. Встает вопрос, как определить $f(X)$.

Математика предоставляет нам возможность найти функцию, отражающую искомую линию регрессии с любой степенью приближения. Это можно сделать, например, используя многочлены произвольной степени m:

$$g(X,\beta) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_m X^m$$

(β_0, β_1, β_2, ..., β_m – некоторые параметры, выборочные оценки которых надо получить). Однако найденная функция, вообще говоря, будет очень сложной и вряд ли с ее помощью мы сможем практически осуществлять прогноз, т.е. достигнем основной цели построения регрессионных моделей. Причины такой непригодности сложных формул частично сходны с теми, что были обсуждены нами в п.2.5.3.2 при рассмотрении третьей причины останова алгоритма THAID, слишком сложные формулы мы в силу своей психологической специфики не можем воспринимать как закономерность (п.1.4 части I).

Чтобы избежать чрезмерной сложности искомой закономерности, обычно выбирают какое-либо семейство кривых, выражающихся сравнительно простыми формулами, и именно среди них с помощью метода наименьших квадратов ищут ту, которая как можно более близко подходит ко всем данным точкам. Чаще всего в качестве такого семейства используют совокупность прямых линий. Как известно, все такие линии выражаются формулами вида

$$g(X,\beta) = \beta_0 + \beta_1 X$$

где β_1 говорит о величине угла наклона прямой к оси X, а β_0 – о сдвиге этой прямой вдоль оси Y. Соответствующий вариант регрессионного анализа называется *линейным*. Он используется

практически чаще всего. Отвечающая ему техника хорошо известна. Выборочные оценки коэффициентов линейного уравнения регрессии находятся с помощью описанного выше метода наименьших квадратов.

В данном случае (10) превращается в соотношение

$$\sum (Y - \beta_0 - \beta_1 X)^2 \in \min$$

Далее мы, условно говоря, как бы «перебираем» все возможные прямые (точнее, все возможные пары чисел β_0 и β_1) и находим ту прямую, для которой наша сумма будет минимальной. Конечно, в действительности перебрать все прямые также невозможно (как известно, совокупность всех действительных чисел нельзя даже «пересчитать» с помощью бесконечного ряда натуральных чисел), параметры искомой прямой ищутся с помощью производных: находим производную от нашей суммы по β_0 и β_1 и ищем те их значения, которые обращают производную в нуль. Получаем известные аналитические выражения для этих коэффициентов (напомним, что латинскими буквами обозначаются выборочные оценки одноименных генеральных параметров):

$$b_0 = \overline{Y} - b_1 \overline{X} = r \frac{S_Y}{S_X}$$

$$b_1 = \sum \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{(X_i - \overline{X})^2},$$

где r – коэффициент корреляции между X и Y; S_Y и S_X – выборочные оценки средних квадратических отклонений соответствующих признаков; суммирование, как и выше, осуществляется по всем объектам.

В идеале точка с координатами $(X, \beta_0 + \beta_1 X)$ должна лежать на линии регрессии. В соответствии с упомянутыми выше традиционными предположениями, это означает справедливость картины, отраженной на рисунке 27.

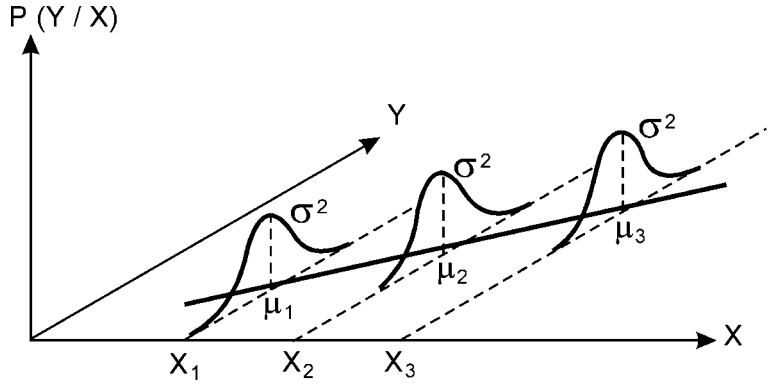

Рис.27. Статистические предположения, лежащие в основе линейного регрессионного анализа

Условные распределения Y нормальны. Их математические ожидания лежат на прямой линии, дисперсии равны

Другими словами, мы предполагаем, что каждому значению независимой переменной X отвечают нормальные гомоскедастичные условные распределения Y, математические ожидания которых принадлежат рассматриваемой прямой. Это предположение эквивалентно следующему соотношению:

$$Y_i = \beta_0 + \beta_1 X_i + e_i,$$

означающему, что каждое наблюдаемое значение Y_i есть сумма некой фиксированной величины $\beta_0 + \beta_1 X$, обусловленной линией регрессии, и случайной величины e_i, обусловленной естественной вариацией значений Y вокруг линии регрессии. При каждом значении независимой переменной X вариация Y имеет тот же характер, что и вариация e_i. Отсюда ясно, что все e_i имеют нормальные распределения с нулевыми математическими ожиданиями и равными дисперсиями σ^2. Важность случайных величин e_i заключается в том, что они представляют собой главный источник ошибок при попытке предсказать Y по значению X. В рамках регрессионного анализа разработаны способы оценки величин e_i.

На практике чаще всего пользуются именно линейными регрессионными моделями. Однако при их использовании необходимо

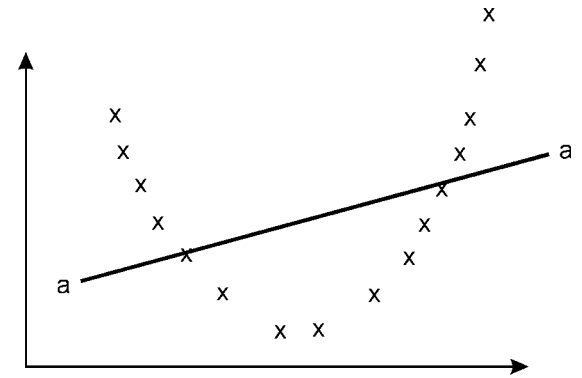

Рис.28. Пример криволинейной линии регрессии между двумя переменными

Несоответствие ей прямой «aa»

учитывать, что идеальная картина, изображенная на рисунке 27 – это лишь наше пожелание. Наилучшая прямая среди всех возможных прямых может быть весьма плохим приближением к реальности. Скажем, если наши крестики расположены так, как это отражено на рисунке 28, то любая прямая (например, «aa») здесь даст очень плохое приближение.

В данном случае надо бы вместо прямых линий для поиска использовать семейство подходящих кривых квадратных трехчленов вида

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Используя же технику линейного регрессионного анализа, и тем самым направляя свою энергию на поиск лучшей прямой, приближающей нашу совокупность точек, мы рискуем никогда не узнать, что в действительности имели дело с линией регрессии, являющейся параболой. Правда, тут необходимо отметить два момента.

Во-первых, для двумерного случая, который мы пока рассматриваем, такое вряд ли случится, поскольку перед нами – наглядная плоскостная картина, глядя на которую всегда можно определить, прямая ли линия соответствует изучаемому множеству

точек или парабола. В случае же многомерного регрессионного анализа, который мы коротко рассмотрим ниже, такой просчет вполне возможен.

Во-вторых, в регрессионном анализе существуют достаточно разработанные подходы к построению регрессионных кривых нелинейного вида. Имеются критерии линейности и рекомендации по выбору степени аппроксимирующего многочлена.

О нелинейных моделях коротко мы еще вспомним ниже (см. п.2.6.5). Пока же коротко рассмотрим многомерный случай, т.е. такую ситуацию, когда имеется много независимых переменных X_1, X_2, ... , X_n ($n >> 1$). Все сказанное выше справедливо и для рассматриваемой ситуации. Отличие состоит только в том, что здесь линейная регрессионная модель имеет вид не прямой линии, а так называемой гиперплоскости:

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + ... + a_n \cdot X_n \ .$$

Здесь необходимо два слова сказать об интерпретации только что выписанного уравнения (в соответствии с общепринятой терминологией, слева пишется просто Y, а не условное среднее $\overline{Y}_{x_1 x_1 \ldots x_n}$, и найденное с помощью техники регрессионного анализа соотношение называется уравнением, хотя этот термин и употребляется не в том смысле, в каком его используют в школе; a_0 называется свободным членом уравнения). Однако прежде сделаем некоторые замечания о единицах измерения рассматриваемых признаков. Интуитивно ясно, что уравнение регрессии будет более ясным с точки зрения его содержательной интерпретации, если все эти единицы будут одинаковыми. Для этого обычно осуществляют так называемую стандартизацию всех значений каждого признака: вычитают из каждого такого значения среднее арифметическое признака (точнее, здесь речь должна идти о математическом ожидании, за неимением которого мы используем его выборочную оценку – среднее арифметическое) и делят полученную разность на его же дисперсию (и снова вместо генеральной дисперсии мы вынуждены пользоваться ее выборочной оценкой). Рассмотрим для примера признак X_2. Если X_2^i – некоторое (i-е) его значение, $\overline{X}_2$ и σ_x – соответственно, отвечающие ему среднее арифметическое и дисперсия, то указанная нормировка будет означать следующее преобразование значения X_2^i:

$$X_2^i \to \frac{X_2^i - \overline{X}_2}{\sigma_x} \ .$$

Нетрудно видеть, что среднее значение нормированного признака будут равно нулю, а дисперсия – единице. Далее будем считать, что описанная нормировка для всех рассматриваемых признаков произведена и что тем саммым снята проблема несравнимости их значений из-за «разномасштабности». Обозначения признаков оставим прежними.

Интерпретация коэффициентов очевидна. Если, скажем, значение признака X_2 изменится на единицу, то значение Y изменится на a_2. Поэтому a_2 можно интерпретировать как величину приращения Y, получаемого за счет увеличения признака X_2 на единицу.

В заключении обсуждения вопроса о классическом регрессионном анализе заметим, что указанная «прозрачная» интерпретация может «затуманиться» в том случае, если наши предикторы связаны друг с другом. Причина тоже довольно очевидна. Поясним это.

Предположим, что X_2 связан с X_5, и мы хотим узнать, на сколько изменится Y при увеличении X_2 на единицу. Рассуждать так же, как выше, мы не можем: увеличение X_2 неумолимо приведет к увеличению (или уменьшению) X_5, и поэтому изменение Y будет обусловлено изменением не только X_2, но и X_5. На сколько изменится X_5, вообще говоря, неизвестно. Чтобы ответить на этот вопрос, нужно подробнее изучить форму зависимости между X_2 и X_5. А это – самостоятельная и, возможно, сложная задача. Без ее решения вопрос о величине изменения Y остается открытым. И в любом случае это изменение, вообще говоря, не будет равно a_2.

В силу сказанного, будем стремиться к тому, чтобы избегать включения в уравнение регрессии заведомо связанных друг с другом предикторов.

Описание идей регрессионного анализа можно найти в [Мостеллер, Тьюки, 1982; Паниотто, Максименко, 1982; Статистические методы..., 1979].

Теперь перейдем к рассмотрению вопроса о возможности использования техники линейного регрессионного анализа к номинальным данным.

2.6.3. Дихотомизация номинальных данных. Обоснование допустимости применения к полученным дихотомическим данным любых «количественных» методов

Конечно, использовать регрессионную технику для анализа «чисел», являющихся метками, отвечающих некоторой номинальной шкале, бессмысленно (считаем это интуитивно ясным, хотя можно было бы доказать такое утверждение строго, используя понятие адекватности математического метода из теории измерений [Толстова, 1998]). Для того, чтобы на основе информации, полученной по номинальной шкале, можно было построить уравнение регрессии, эту информацию необходимо преобразовать. Соответствующее преобразование носит название дихотомизации номинальных данных. Этот подход применяется очень широко, поскольку его использование как бы «открывает дверь» для применения подавляющего большинства «количественных» методов с целью анализа номинальных данных. Опишем суть преобразования.

Вместо каждого номинального признака, принимающего «k» значений, вводим «k» новых дихотомических (т.е. принимающих два значения, будем обозначать эти значения 0 и 1). Надеемся, что то, как это делается, станет ясным из следующего примера.

Предположим, что рассматриваемый номинальный признак X – это национальность и что в соответствующем закрытом вопросе анкеты фигурируют три национальности: русский, грузин и чукча. Каждой из этих альтернатив поставим свой дихотомический признак, задаваемый следующим правилом (напомним, что задать признак – значит задать правило приписывания отвечающих ему значений каждому респонденту):

$$\text{русский} \rightarrow X_1 = \begin{cases} 1, \text{ если рассматриваемый респондент – русский;} \\ 0, \text{ если рассматриваемый респондент – не русский} \\ \text{(кто именно – грузин или чукча – безразлично)} \end{cases}$$

$$\text{грузин} \rightarrow X_2 = \begin{cases} 1, \text{ если рассматриваемый респондент – грузин} \\ 0, \text{ если рассматриваемый респондент – не грузин} \\ \text{(кто именно – русский или чукча – безразлично)} \end{cases}$$

$$\text{чукча} \rightarrow X_3 = \begin{cases} 1, \text{ если рассматриваемый респондент – чукча} \\ 0, \text{ если рассматриваемый респондент – не чукча} \\ \text{(кто именно – русский или грузин – безразлично)} \end{cases}$$

Применение регрессионной техники к преобразованным номинальным данным называется номинальным регрессионным анализом. При этом Y может быть не только номинальным, но и интервальным. Поясним подробнее, что именно при реализации соответствующего подхода происходит с зависимой и независимыми переменными. Предположим, что мы хотим изучить связь вида

$$Y = f(X),$$

где X – скажем, та же национальность (предусматривающая, как и выше, три варианта ответов), а Y – профессия. Вместо признака X в уравнение необходимо вставить три новых предиктора – X_1, X_2, X_3, описанные выше. Однако здесь имеется один нюанс. В конце п.2.6.1 мы отмечали нежелательность включения в регрессионную модель таких предикторов, которые заведомо связаны друг с другом. А относительно наших X_1, X_2, X_3 такая связь как раз имеет место. Покажем это.

Нетрудно видеть, что если мы знаем значения двух из трех рассматриваемых предикторов, то значение третьего определяется автоматически. Мы можем не спрашивать респондента, какая у него национальность, а сами определить ее, если знаем, какие значения для него имеют признаки X_1 и X_2. Это демонстрируется приведенной ниже таблицей 28.

Если человек – не русский и не грузин, то он – чукча; если он русский, а не грузин, то он и не чукча; если же он не русский, но грузин, то он тоже не чукча; быть же одновременно и русским, и грузином он не может.

Таблица 28

Иллюстрация зависимости друг от друга признаков, являющихся результатом дихотомизации одной номинальной переменной

Заданные значения признаков		Теоретически определяемое значение признака
X_1	X_2	X_3
0	0	1
1	0	0
0	1	0

Поэтому во избежание недоразумений, могущих возникнуть при интерпретации результатов регрессионного анализа, желательно не включать в уравнение все три дихотомические переменные. Именно так обычно и поступают. Один дихотомический признак как бы отбрасывают (ниже мы увидим, что это отбрасывание в содержательном плане является фиктивным: в процессе интерпретации коэффициентов найденного уравнения сведения об отброшенном признаке будут присутствовать). Таким образом, число аргументов искомого уравнения будет на единицу меньше, чем число альтернатив в рассматриваемом номинальном признаке. В нашем случае вместо трех предикторов мы включаем в уравнение только два. Ниже будем считать, что мы отбросили X_3.

Теперь рассмотрим ситуацию с зависимой переменной Y. Она так же, как и X превращается в несколько дихотомических признаков. Пусть, например, в нашей анкете предусмотрено три варианта ответа – учитель, торговец, дворник. Тогда вместо Y возникают три следующие дихотомические признака:

$$Y_1 = \begin{cases} 1, \text{ если респондент – учитель,} \\ 0, \text{ если респондент – не учитель;} \end{cases}$$

$$Y_2 = \begin{cases} 1, \text{ если респондент – торговец,} \\ 0, \text{ если респондент – не торговец;} \end{cases}$$

$$Y_3 = \begin{cases} 1, \text{ если респондент – дворник,} \\ 0, \text{ если респондент – не дворник.} \end{cases}$$

Встает вопрос: какой из этих новых Y-ков необходимо взять в качестве независимой переменной искомого уравнения регрессии (ясно, что использование сразу нескольких зависимых переменных бессмысленно). Выход довольно очевиден: надо строить три уравнения регрессии, каждое из которых отвечает своему Y_i.

Итак, задача сводится к построению следующей системы уравнений регрессии (термин «система» здесь употреблен не случайно: уравнения взаимосвязаны и содержательно дополняют друг друга):

$$Y_1 = f_1 (X_1, X_2),$$

$$Y_2 = f_2 (X_1, X_2),$$

$$Y_3 = f_3 (X_1, X_2).$$

Как мы уже отмечали, техника нахождения конкретного вида каждого уравнения традиционна – это техника «числового» регрессионного анализа.

Попытаемся ответить на вопрос о том, почему такая подмена возможна, т.е. почему к числам, полученным по произвольной номинальной шкале, применять регрессионную технику (равно как и любой другой «количественный» метод) нельзя, а к отвечающим номинальной же шкале 0 и 1 – можно (и это «разрешение» тоже касается не только регрессионного анализа). Напомним, что аналогичный вопрос применительно к вычислению среднего арифметического уже рассматривался нами в п.1.2. В настоящем и следующем параграфе мы обсудим его в более общей постановке.

Во-первых, с формальной точки зрения упомянутую дихотомическую номинальную шкалу можно рассматривать как частный случай интервальной. Здесь мы имеем дело только с одним интервалом – между 0 и 1. И представляется вполне допустимой истинность утверждения: за равными числовыми интервалами стоят некоторые реальные равные эмпирические разности между объектами.

Во-вторых, допустимость применения количественного метода к дихотомическим данным опирается на то, что, как оказывается, многие известные математические статистики, будучи

вычисленными для таких данных, как правило, оказывается возможным проинтерпретировать вполне разумным образом, чего отнюдь нельзя сказать об интерпретации соответствующих показателей, вычисленных для многозначных номинальных шкал.

Пример вычисления среднего арифметического для пола респондента, приведенный в разделе 1, подтверждает это (отметим, однако, что полу отвечает естественная дихотомия, а не искусственная, как в рассмотренных выше ситуациях; иногда естественные и искусственные дихотомии противопоставляют друг другу; однако для нас это не актуально). Демонстрация того, что осмысленная интерпретация возможна и для найденных рассматриваемым образом коэффициентов уравнения регрессии, будет осуществлена в п.2.6.4.

Последнее обстоятельство, на котором нам хотелось бы остановиться в данном параграфе, состоит в том, что, как оказывается, задача применения традиционной регрессионной техники остается осмысленной и для того случая, когда Y измеряется по интервальной шкале. Специфика такой ситуации проявляется в интерпретации результатов регрессионного анализа. Ниже на этом мы также остановимся.

2.6.4. Общий вид линейных регрессионных уравнений с номинальными переменными. Их интерпретация

Итак, предположим, что у нас имеются некоторые номинальные признаки Y (зависимый; пока, до обсуждения некоторых вопросов, связанных с интерпретацией результатов регрессионного анализа, будем считать этот признак номинальным) и X_1 , X_2 ,..., X_n (независимые). Пусть Y принимает k значений, а каждый признак X_i – l_i значений. Предположим также, что осуществлена дихотомизация исходных данных, в результате чего независимый признак «превращен» в дихотомические признаки Y_1, Y_2, ... , Y_k , а каждый признак X^i – в дихотомические X_1^i , X_2^i, ..., $X_{1_i}^i$. Будем полагать, что в качестве «отбрасываемого признака фигурирует последний признак из каждого только что приведенного набора.

Применение техники номинального регрессионного анализа к такого рода данным означат расчет k уравнений вида:

$$Y_1 = f_1(X_1, X_2,...,X_n) = f_1(X_1^1, X_2^1,..., X_{1_{1-1}}^1, X_1^2, X_2^2, ..., X_{1_{2-1}}^2, ... , X_1^n, X_2^n, ..., X_{1_{n-1}}^n)$$

$$Y_2 = f_2(X_1, X_2,...,X_n) = f_2(X_1^1, X_2^1,..., X_{1_{1-1}}^1, X_1^2, X_2^2, ..., X_{1_{2-1}}^2, ..., X_1^n, X_2^n, ..., X_{1_{n-1}}^n)$$

.....

$$Y_k = f_k(X_1, X_2,...,X_n) = f_k(\underbrace{X_1^1, X_2^1,..., X_{1_{1-1}}^1}_{\text{отвечают } X^1}, \underbrace{X_1^2, X_2^2, ..., X_{1_{2-1}}^2}_{\text{отвечают } X^2},..., \underbrace{X_1^n, X_2^n, ..., X_{1_{n-1}}^n}_{\text{отвечают } X^n})$$

Не хотим далее «мучить» читателя индексами и поэтому все дальнейшие рассуждения будем вести в предположении, что рассматривается только одна градация зависимого признака с отвечающей ей дихотомической переменной Y и один принимающий три значения независимый признак X с отвечающими ему дихотомическими переменными X_1, X_2, X_3. Надеемся, что необходимые обобщения читатель сделает самостоятельно.

Таким образом, будем полагать, что искомая зависимость имеет вид:

$$Y = f(X_1, X_2) = a_0 + a_1 \text{ w } X_1 + a_2 \text{ w } X_2 \qquad (11)$$

Например, предположим, что Y, X_1, X_2 – это дихотомические переменные, отвечающие, соответственно, свойствам «быть торговцем», «быть русским» и «быть грузином» (напомним, что дихотомическую переменную, отвечающую свойству «быть чукчей», мы при построении уравнения отбрасываем). Процесс поиска подобной зависимости состоит в реализации техники линейного регрессионного анализа.

Коэффициенты уравнения регрессии, найденные по всем правилам классического регрессионного анализа, выражаются довольно сложными формулами, включающими в себя такие (вроде бы «запретные» для номинальных данных) статистики, как среднее арифметическое, дисперсия, частные коэффициенты

корреляции и т.д., Однако, как мы уже упоминали, их оказывается возможным проинтерпретировать вполне разумным , понятным любому социологу, способом – как некоторые условные частоты. Опишем эту интерпретацию.

Сначала проинтерпретируем коэффициент a_0 (свободный член уравнения (11)). В силу самой сути уравнения регрессии, подставив в него произвольные значения независимых переменных X_1, X_2, слева от знака равенства мы получим среднее значение Y, которое отвечает совокупности респондентов с рассматриваемыми значениями предикторов. Рассмотрим только тех людей, которым соответствует отброшенная нами национальность, – чукчей. Ясно, что для них $X_1 = X_2 = 0$. Подставив эти значения в уравнение регрессии, получим соотношение

$$Y = a_0 .$$

Таким образом, интерпретируемый коэффициент a_0 равен среднему арифметическому значению зависимой переменной для отброшенной категории респондентов, в данном случае – для чукчей. Если бы Y был *интервальной* переменной, то тем самым интерпретация свободного члена уравнения регрессии была бы окончена. Но наш Y – дихотомическая переменная, отвечающая свойству «быть торговцем». В соответствии с описанной выше интерпретацией среднего арифметического значения дихотомического признака, смысл a_0 сводится к тому, что это – доля чукчей, работающих торговцами (говоря формально – доля отброшенной категории респондентов, обладающих единичным значением зависимого признака).

Перейдем к интерпретации коэффициента a_1 из уравнения (11). Рассмотрим только русских. Нетрудно видеть, что для них $X_1 = 1$ и $X_2 = 0$. Подставим эти значения в уравнение. Получим соотношение:

$$Y = a_0 + a_1.$$

Учитывая осуществленную выше интерпретацию свободного члена уравнения, применительно к нашему примеру, можно сказать, что a_1 – это тот «довесок», который надо прибавить к доле

чукчей, являющихся торговцами, чтобы получить долю русских, занимающихся этим делом. Аналогична интерпретация $a_{2:}$: это та величина, которую надо прибавить к доле торговцев среди чукчей, чтобы получить аналогичную долю среди грузин. Приведем пример.

Пусть уравнение, найденное с помощью линейного регрессионного анализа имеет вид:

$$Y = 0{,}3 – 0{,}1\ X_1 + 0{,}6\ X_2 \qquad (12)$$

Его коэффициенты можно интерпретировать как условные частоты: доля торговцев среди чукчей равна 0,3, среди русских – (0,3 + (– 0,1)) = 0,2, а среди грузин – (0,3 + 0,6) = 0,9.

Чтобы еще более стал ясен смысл коэффициентов уравнения регрессии, рассмотрим, во что это уравнение превращается в случае изучения двух дихотомических признаков. Приведем пример из [Типология и классификация..., 1982, с.260-266]. Пусть X – семейное положение (два значения: X_1 – женат, X_2 – неженат), Y – посещение кинотеатра (Y_1 – посещает, Y_2 – не посещает; здесь мы отвлекаемся от точного смысла этих слов: означает ли выражение «не посещает» то, что респондент никогда не ходил в кино, или же что он не был там в течение последних 5-ти лет и т.д.).

Пусть таблица сопряженности,, отвечающая нашим признакам, имеет вид:

Таблица 29

Общий вид четырехклеточной таблицы сопряженности

Y	X		Итого
	X_1	X_2	
Y_1	a	b	a + b
Y_2	c	d	c + d
Итого	a + c	b + d	a + b + c + d

Найдем коэффициенты уравнения регрессии вида

$$Y = \alpha + \beta X.$$

В соответствии с нашими правилами, они равны:
α = b / b+d (доля посещающих кинотеатр среди неженатых);

$\beta = \frac{a}{a+c} - \frac{b}{b+d}$ (тот «довесок», который надо прибавить к доле посещающих кинотеатр среди неженатых, чтобы получить аналогичную долю среди женатых (последняя равна $\frac{a}{a+c}$) .

Приведем соответствующий цифровой пример. Пусть конкретная матрица имеет вид:

Таблица 30

Пример четырехклеточной таблицы сопряженности

Y	X		Итого
	X_1	X_2	
Y_1	48	38	86
Y_2	2	12	14
Итого	50	50	100

Тогда верны соотношения:

$$\alpha = \frac{b}{b+d} = \frac{38}{50} = 0{,}76\,;\ \beta = \frac{a}{a+c} - \frac{b}{b+d} =$$

$$= 0{,}96 - 0{,}76 = 0{,}2$$

$Y = 0{,}76 + 0{,}2\,X.$

Нетрудно увидеть связь между номинальным регрессионным и детерминационным анализом. Действительно, в соответствии с последним, $I(X_2 \epsilon Y_1) = P(Y_1/X_2) = \frac{38}{50} = \alpha$. В то же время

$I(X_1 \epsilon Y_1) = P(Y_1/X_1) = \frac{48}{50}$ и поэтому $\beta = I(X_1 \epsilon Y_1) - I(X_2 \epsilon Y_1)$.

Итак, все коэффициенты рассматриваемого уравнения регрессии интерпретируются через некоторые условные частоты. Встает вопрос: надо ли использовать сложную технику

регрессионного анализа для того, чтобы получить результаты, получаемые обычно социологом более простым путем (путем прямого расчета многомерных частотных таблиц)? Покажем, что такая постановка вопроса неправомерна: регрессионный анализ нельзя свести только к получению условных частот. Уравнение регрессии представляет собой систему, свойства которой не сводятся к свойствам отдельных составляющих ее элементов (коэффициентов найденного уравнения). Рассмотрим это обстоятельство подробнее.

2.6.5. Типы задач, решаемых с помощью НРА. Краткие сведения о логит- и пробит-моделях регрессионного анализа

Итак, *первый тип* решающихся с помощью НРА задач – это нахождение определенных условных процентов. Однако, как мы уже заметили, интерпретация результатов регрессионного анализа не сводится к интерпретации отдельных коэффициентов уравнения регрессии. Выше, в начале нашего рассмотрения этого подхода, мы говорили о том, что основная цель его использования в любой науке состоит в получении возможности определенного рода прогноза. Попытаемся проинтерпретировать модели номинального регрессионного анализа с соотвествующей точки зрения.

Вернемся к модели общего вида:

$$Y_1 = f_1(X_1, X_2, ..., X_n) =$$
$$= f_1(X_1^1, X_2^1, ..., X_{1_{1-1}}^1, X_1^2, X_2^2, ..., X_{1_{2-1}}^2, ..., X_1^n, X_2^n, ..., X_{1_{n-1}}^n)$$

Сначала предположим, что мы используем линейные модели.

По тому, какие из коэффициентов уравнения регрессии принимают наибольшие значения, можно судить о тех сочетаниях значений независимых признаков, которые в наибольшей мере детерминируют наличие у респондентов единичного значения зависимого. Другими словами, можно осуществлять поиск взаимодействий. Здесь явно просматривается связь с теми задачами, на решение которых направлены рассмотренные выше алгоритмы

типа AID (напомним, более или менее подробно мы рассмотрели алгоритмы THAID и CHAID в п.2.5.3.2 и 2.5.3.3 соответственно). Это – *второй тип* задач. Опишем способы их решения более подробно.

Пусть X_1 – как выше, национальность с градациями (русский, грузин, чукча), X_2 – место проживания с градациями (город, село, кочевье), Y – дихотомическая переменная, отвечающая профессии «торговец». И если при подсчете уравнения линейной номинальной регрессии, к примеру, окажется, что сравнительно большими являются коэффициенты при дихотомических переменных X_2^1 (отвечающей свойству «быть грузином») и X_1^2 (жить в городе), то это будет означать, что именно эти два свойства в совокупности определяют тот или иной уровень доли торговцев в изучаемой группе респондентов. Представляется очевидным сходство этих выводов с теми, которые позволяют получать алгоритмы THAID и CHAID.

Еще более надежными станут выводы подобного рода, если мы будем использовать нелинейные модели. Сразу подчеркнем, что в номинальном регрессионном анализе гораздо легче решается проблема выбора модели, чем в «числовом» варианте этого анализа. Так, здесь резко сокращается круг тех многочленов, среди которых имеет смысл искать интересующие нас закономерности. В частности, ни к чему вставлять в искомое уравнение степени рассматриваемых переменных, поскольку для любого дихотомического признака любая его степень равна самому признаку (так как $0^2 = 0$, $1^2 = 1$). А вот произведения переменных имеет смысл включить. Эти произведения отвечают тем самым взаимодействиям, о которых шла речь выше.

Например, если доля торговцев среди изучаемых респондентов определяется долей горожан-грузин, то мы, несомненно, это выявим путем включения в уравнения произведения вида X_2^1 w X_1^2 (обозначения – как выше).

Ясно, что произведения трех дихотомических переменных будут отвечать «трехмерным» взаимодействиям и т.д.

Третий тип задач связан с возможностью осуществлять прогноз несколько иного вида. Поясним это на примере. Вернемся к

соотношению (12). В силу его очевидных арифметических свойств, можно сказать, что коэффициенты –0,1 и 0,6, соответственно, означают вклад свойств «быть русским» (X_1) и «быть грузином» (X_2) в долю торговцев (Y) среди респондентов изучаемой совокупности. Однако проинтерпретировать смысл этого вклада трудно при дихотомических переменных. Поэтому часто прибегают к следующим рассуждениям, опирающимся на довольно сильные модельные предположения. Полагают, что указанное уравнение справедливо не только для того случая, когда X_1 и X_2 – дихотомические переменные, характеризующие отдельных респондентов, но для такой ситуации, когда в качестве единиц наблюдения фигурируют группы людей, а X_1 и X_2 – доли, соответственно, русских и грузин в этих группах. В таком случае смысл уравнения становится ясным: если доля русских увеличивается в группе, скажем, на 10%, то доля торговцев увеличивается на (–0,1) w 10% = –1% (т.е. уменьшается на 1%). Если же доля грузин в совокупности увеличивается на 10%, то доля торговцев увеличивается на (0,6) w 10 % = 6%.

Заметим, что класс решаемых с помощью техники номинального регрессионного анализа задач может быть расширен за счет использования приемов, широко применяющихся во всем мире при анализе статистического материала, но не рассмотренных в настоящем учебнике. Мы имеем в виду так называемые обобщенные линейные модели (Generalized Linear Model, GLM), в частности, логистическую регрессию, использование так называемых логит-моделей. Коротко опишем суть подхода, уделив особое внимание тому случаю, когда Y – дихотомическая номинальная переменная. То, о чем пойдет речь, можно найти в работах [Agresti, 1996, ch.4; Demaris, 1992, ch.4; Menard, 1995].

Напомним, что линейное регрессионное уравнение чаще всего имеет следующий вид:

$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k .$$

Левая часть этого уравнения обычно связывается со *случайной компонентой* рассматриваемой линейной модели. Эта компонента говорит о том, что объясняемая переменная Y является случайной

величиной с математическим ожиданием μ. О правой части говорят как о *систематической компоненте* линейной модели. При этом понятие линейности зачастую расширяется: допускается, что одни x_i могут выражаться через другие. Например, наличие переменной вида $x_3 = x_1 x_2$ говорит о взаимодействии между x_1 и x_2 в процессе их воздействия на Y. Наличие переменной вида $x_3 = x_1^2$ свидетельствует о криволинейности воздействия x_1 на Y.

Очень важным элементом рассматриваемой модели является форма связи между случайной и систематической компонентами модели. Выше мы говорили о сложности выбора этой формы. Но при этом полагали, что разные виды зависимости можно отразить с помощью преобразования правой части модели. Однако имеет смысл преобразовывать и левую часть. Так, в литературе по анализу данных принято называть *связующей функцией* (link function) такую функцию g, для которой справедливо соотношение

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

Если g – тождественная функция ($g(\mu) = \mu$, identity link), то только что написанное соотношение превращается в обычную регрессию. Если же g – это логарифм (log link), то получаем то, что называется *логлинейной моделью*:

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

Преимущество использования логлинейной модели заключается в том, что она дает возможность свести изучение сложных взаимодействий между независимыми переменными (т.е. подбор таких произведений x-ов, которые делают адекватной реальности используемую модель; выше мы говорили о важности и трудности решения этой задачи) к поиску коэффициентов линейной зависимости (поскольку логарифм произведения равен сумме логарифмов).

Особую важность имеет так называемая логит-связь (logit link), когда функция g является функцией вида:

$$g(\mu) = \log \frac{\mu}{1-\mu}$$

Обобщенная линейная модель при использовании такой связи называется *логит-моделью* (logit model). Эта модель играет большую роль в тех случаях, когда Y – дихотомическая переменная. Используя введенные выше обозначения (p – доля единичных значений Y, а $q = (1 - p)$ – доля нулевых значений того же признака) можно сказать, что здесь

$$g(\mu) = \log \frac{p}{q} .$$

Другими словами, функция g является логарифмом отношения преобладания. Ниже для простоты будем предполагать, что у нас только один признак X. Уравнение вида

$$\log \frac{p(X)}{1-p(X)} = \alpha + \beta X$$

называется *логистической регрессионной функцией*. Важность ее изучения представляется очевидной (скажем, для приведенного в предыдущих параграфах примера она позволяет выявить причины изменения соотношения читающих и нечитающих данную газету).

Не менее очевидной является важность изучения и так называемой *линейной вероятностной модели*

$$p(X) = \alpha + \beta X$$

(применительно к тому же примеру, здесь речь идет об изменении доли читающих газету). Заметим, что, когда независимых переменных много, подобного рода уравнения совпадают с теми, которые обычно связываются с логлинейным анализом (там в качестве значений независимой переменной выступают частоты многомерной таблицы сопряженности).

Описанные модели являются очень полезными для социолога. Для интерпретации полученных с их помощью результатов можно использовать описанные в п.2.6.4 приемы. Отличие будет состоять в трактовке того, что стоит в левой части найденного регрессионного уравнения. Эта трактовка определяется тем, что было только что сказано нами. Ясно, что использование упомянутых моделей расширяет круг решаемых с помощью НРА задач.