Министерство образования и науки Российской Федерации
Федеральное агентство по образованию
Ярославский государственный университет им. П.Г. Демидова

О.В. Епархина

Математические методы обработки и анализа социологических данных

Учебное пособие

Рекомендовано
Научно-методическим советом университета
для студентов специальности Социология

Ярославль 2007

УДК 316:303.7
ББК С5в6я73
Е 65

Рекомендовано
Редакционно-издательским советом университета
в качестве учебного издания. План 2007 года

Рецензенты:
доктор технических наук, профессор А.А. Мурашов;
кафедра политологии ЯГТУ

Епархина, О.В. Математические методы обработки и анализа
Е 65 социологических данных : учеб. пособие / О.В. Епархина; Яросл. гос. ун-т. – Ярославль : ЯрГУ, 2007. – 132 с.
ISBN 978-5-8397-0527-2

В пособии даны теоретические и практические аспекты использования математики в социологии, описаны конкретные методы анализа социологических данных. Представлены алгоритмы поиска связей между номинальными признаками (коэффициенты связи, многомерные отношения преобладания, сочетания независимых предикторов и т.п.).

Учебное пособие предназначено для студентов, обучающихся по специальности 020300 Социология (блок ЕН) очной формы обучения.

УДК 316:303.7
ББК С5в6я73

ISBN 978-5-8397-0527-2

Введение

Настоящая работа является учебным пособием по курсу «Математические методы обработки и анализа социологических данных», читаемому автором для студентов-социологов. Практически в каждом учебном заведении, готовящем социологов, преподается эта дисциплина, что обусловлено возрастанием значения математического аппарата в аналитической работе с данными как в рамках используемых ПК статпакетов, так и в традиционных формах.

В пособии разъясняется специфика анализа социологических данных и показана необходимость использования математики в социологии. В нем содержится описание методов анализа данных, характерных именно для социологии: методов т.н. описательной (дескриптивной) статистики – выборочного представления одномерного вероятностного распределения и расчета его основных параметров (мер средней тенденции и показателей разброса), простейших методов изучения связей между номинальными признаками, а также рассматриваются коэффициенты связи для ранговых признаков, элементы дисперсионного и факторного анализа и т.п.

В отечественной литературе еще в 1970 – 1980-х гг. было представлено много работ, предназначенных для изучения социологами математических методов, использующихся в решении социологических задач. Однако большинство этих работ изданы давно, недоступны студентам, некоторые методы описаны недостаточно подробно.

Приоритет в разработке рассматриваемых методов принадлежит западным ученым. Прежде всего речь идет о

работах А. Агрести, ставших классикой на Западе (логлинейные, логит-, пробит-модели, ряд моделей логистической регрессии, алгоритмы анализа отношений преобладания и т.д.), о монографиях Г. Аптона. В основу данного пособия положены исследования этих крупнейших специалистов, а также Ю.Н. Толстовой.

Предполагается, что студент, приступивший к изучению данного пособия, имеет элементарные знания из курсов по общей социологии, методике социологических исследований, математической статистике.

Тема 1
Общие аспекты применения математических методов в социологическом анализе

1.1. Статистические закономерности в анализе социологической информации

В науке принято выделять две основные формы закономерной связи явлений, отличающиеся по характеру вытекающих из них предсказаний: динамические и статистические закономерности[1]. В законах динамического типа предсказание имеет точный, определенный, однозначный вид; в статистических же законах предсказание носит не достоверный, а лишь вероятностный характер. Нас интересуют в основном статистические закономерности (закономерности «в среднем»).

Статистическая закономерность возникает как результат взаимодействия большого числа элементов, составляющих совокупность, и характеризует не столько поведение отдельного элемента совокупности, сколько всю совокупность в целом. Она адекватно описывает массовые явления случайного характера, а именно такого рода явления и изучает обычно социолог[2]. **Анализ данных с помощью математических методов позволяет выявлять статистические закономерности.**

Но для социологии важен и поиск **динамических закономерностей**: в результате строятся модели мобильности групп в социальных системах, модели процессов межличностного влияния и внутриличностных конфликтов, модели подражательного поведения и т.д.[3]

---

[1] Философский энциклопедический словарь. М.: Наука. 1983. С. 653

[2] Подробнее см.: Толстова Ю.Н. Измерение в социологии. М.: Инфра-М, 1998.

[3] Бартоломью Д. Стохастические модели социальных процессов. М.: Финансы и статистика,1985. Гл. 1 – 2.

Кроме того, социолога должны интересовать такие явления, которые **не носят статистического характера**: например, каким образом среди рабочих-металлургов, средний возраст которых равен 30 годам, встречаются отдельные люди старше 60 лет; почему при отсутствии статистической связи между полом выпускника школы и выбором им профессии на социологический факультет поступили практически одни девушки и т. п. явления (некие «переломные» точки системы).

При изучении социальных явлений мы имеем дело не с самой реальностью, а с ее моделью (формализованной приблизительной реальностью), для исследования которой используется математический аппарат.

В исходных данных можно выделить как бы два аспекта:

– множество скрывающихся за данными реальных объектов (отдельных людей, социальных групп, институтов и т.д.) – *содержательный аспект;*

– получающаяся в результате непосредственного сбора данных совокупность отражающих эти объекты формальных конструктов: чисел, текстов и т.п. – *формальный аспект.*

Совокупность априорных представлений социолога, не предполагающих не только абстрагирования от уникальности изучаемых объектов, но и самого вычленения этих объектов, образуют основу *априорной содержательной модели.* А вычленение в реальности объектов связано с формированием и операционализацией понятий, т.е. выбором конкретных объектов измерения и способов сбора данных (часть *концептуальной модели* реальности).

Построение концептуальной модели включает в себя:

– формирование понятий для измерения признаков (каким образом опрашивать людей, задействовать ли способы шкалирования и т.д.). При использовании количественных методов необходимо определить точный набор значений измеряемых признаков, расположение соответствующих вариантов ответов в анкете, структуру вопроса и т.д. Применяя качественные методы – выявить метод кодировки текстов, общие свойства у разных респондентов;

– определение непосредственно измеряемых объектов (построение и корректировка выборки), решение проблем, связан-

ных с реализацией процедуры измерения (учет влияния интервьюера на результат опроса) и т. п.;

– построение эмпирической и математической систем для обеспечения адекватности математического аппарата характеру решаемой социологической задачи.

Реализация выбранных способов сбора данных приводит нас к фрагменту *формальной модели* реальности.

Итак, в процессе интерпретации подлежащих анализу данных мы выделили их содержательный, концептуальный и формальный (математический) аспекты. Они отвечают построению априорной содержательной, концептуальной и формальной модели реальности в процессе измерения. Аналогичные аспекты можно выделить и в понимании искомой закономерности.

В результате работы с данными мы выявим содержательные и формальные закономерности. Формальная закономерность служит лишь статистическим подтверждением правильности нашего предположения о существовании содержательной закономерности.

Между содержательной и формальной закономерностью стоит *концептуальная модель реальности.* Мы вычленяем соотношения, которые называем, к примеру, наличием связи между рассматриваемыми понятиями, – это даст нам основания для выбора конкретного способа анализа данных (формализма).

В итоге мы приходим к формальной (математической) модели изучаемой социальной реальности. Интерпретация этой модели позволяет сделать содержательные выводы, т. е. приводит исследователя к *апостериорной содержательной модели* реальности.

В социологии острота проблемы адекватного соотнесения реальности с ее формальной (математической) моделью объясняется тем, что построение моделей в значительной мере определяется субъективным видением мира социологом и возможностью формализовать явления множеством способов. Так, известно более 100 способов измерения показателей связи между двумя признаками. Каждый из них отражает лишь какую-то одну сторону связи (Пирсоновский коэффициент корреляции, ранговый коэффициент Кендалла, какой-либо из энтропийных коэффициентов связи).

Предположим, что мы хотим изучить влияние социально-экономического положения в стране на воспитание молодежи[4].

Априорная модель. По существу мы уже опираемся на какие-то априорные модельные соображения, когда формулируем проблему именно указанным образом (другой социолог сформулировал бы ее по иному или вообще не увидел бы здесь проблемы). О реальных объектах пока имеем смутное представление: это предположительно либо молодежь, либо дети, либо те, кто их воспитывает (воспитатели детских садов, учителя, деятели культуры, СМИ и т.д.). Именно в их характеристиках (пока неизвестных) проявляется и социально-экономическое положение, и проблемы воспитания. Об отношениях между реальными объектами, условно названных нами содержательной закономерностью, тоже пока известно мало; мы просто предполагаем, что социально-экономическое положение как-то влияет на воспитание молодежи.

Концептуальная модель. Будем рассматривать только учителей (тем самым вычленим изучаемые объекты): выявим, как наша проблема проявляется в их жизни. Выделим некоторые стороны жизни учителей с помощью понятий «материальное положение учителя» и «производительность его труда» (формирование показателей). Будем полагать, что нас интересует причинно-следственное отношение между этими аспектами жизни учителя. Затем мы должны найти способ выражения названных понятий через наблюдаемые признаки, т.е. осуществить их операционализацию. Считаем, что первое понятие хорошо отражается признаком «зарплата учителя», а второе – признаком «средний процент успеваемости в классах». В качестве меры связи может служить коэффициент корреляции Пирсона. Вычислив конкретное значение этой меры (например, 0,8), получаем формальную закономерность, **формальную модель.**

Здесь важна связь между типом шкалы признака (см. тему 4) и коэффициентом связи. Например, существуют коэффициенты связи, рассчитанные на номинальные шкалы (коэффициенты, основанные на критерии χ-квадрат), порядковые шкалы (коэффициенты Спирмена и Кендалла), интервальные шкалы (коэффициент

[4] Толстова Ю.Н. Анализ социологических данных. М.: Научный мир, 2000. С. 14 – 17.

Пирсона). Тип используемых шкал определяется многими обстоятельствами. Например, если материальное положение учителя измеряется его зарплатой, и мы поделим всех учителей на тех, которые получают зарплату, не превышающую стоимость потребительской корзины, и тех, зарплата которых превышает эту границу, шкала будет номинальной дихотомической. Если выделять три группы учителей – (1) обеспеченных ниже потребительской корзины, (2) на уровне потребительской корзины и (3) выше этого уровня, то используем порядковую шкалу. Если считать, что различие между учителями, получающими 3 400 и 3 600 рублей, то же, что и между учителями, получающими 400 и 600 рублей, применяется интервальная шкала. В каждом случае мы определяем свой коэффициент, одновременно выбирая и способ интерпретации результатов измерения связи.

Выбор *коэффициента корреляции Пирсона* предполагает следующую гипотезу: при переходе зарплаты от 400 к 600 рублям эффективность работы учителя в среднем возросла настолько же, насколько в среднем она возросла при переходе от 3 400 к 3 600 рублям (коэффициент, близкий к 1, говорит о наличии содержательной связи). Но повышение зарплаты учителя от 3 400 до 3 600 рублей, действительно, можно интерпретировать как получение учителем возможности регулярно покупать новые книги и, вследствие этого, более эффективно работать. Но данный вывод не распространяется на повышение зарплаты от 400 до 600 рублей: эти зарплаты не могут поднять материальное положение учителя даже на уровень продовольственной корзины. Здесь причина может быть в возрасте, чувстве долга, стереотипах и т. п.

Если использовать какой-либо из *порядковых коэффициентов корреляции*, интерпретации будут другие. Так, если окажется, что люди, живущие в нищете, в среднем хуже работают, чем люди, живущие в бедности, а последние – в среднем хуже, чем те, которые смогли «вылезти» из бедности, у нас будут основания говорить о подтверждении закономерности. Именно такой вывод позволит сделать близость порядковых коэффициентов к 1[5]. Схематично этот пример изображен на рис. 1.

[5] Толстова Ю.Н. Анализ социологических данных. С. 14 – 17.

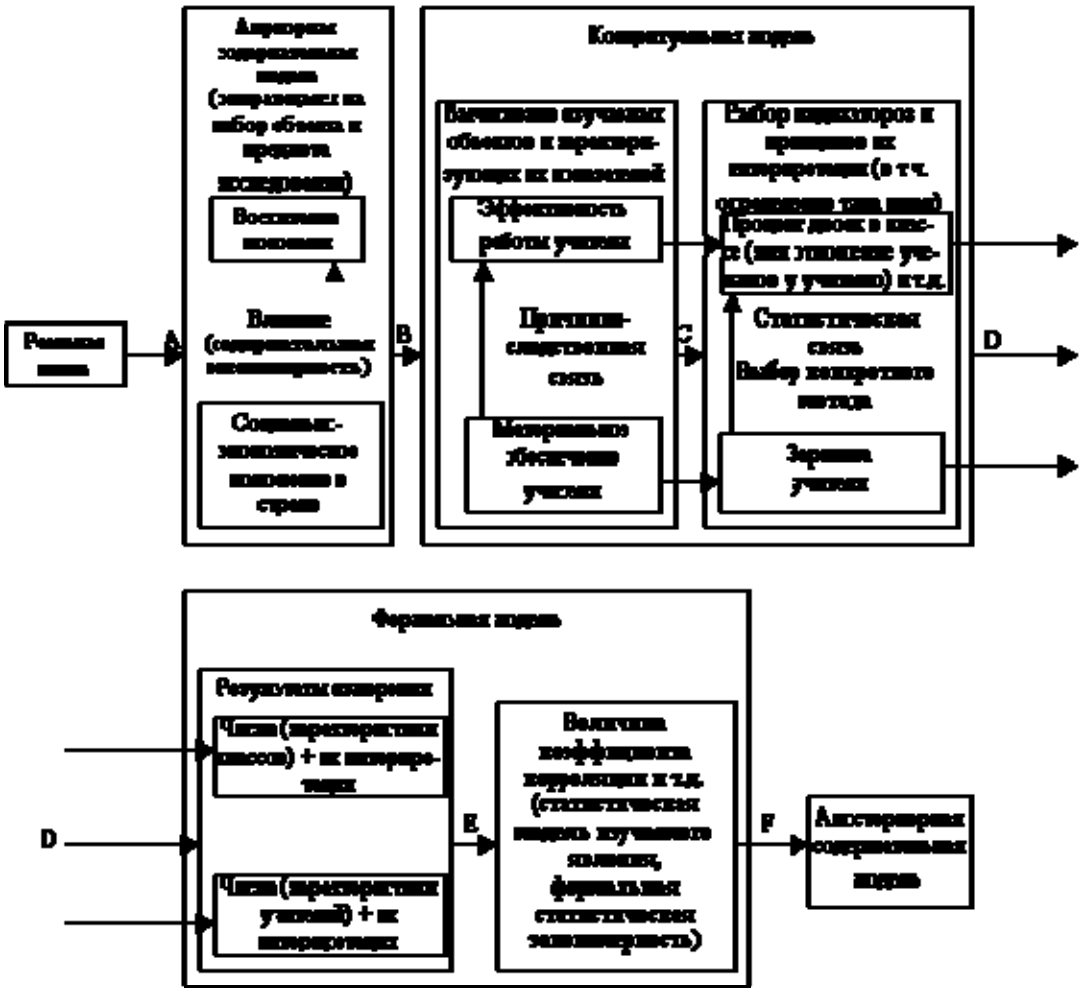


Рис. 1. Формирование и операционализация понятий при анализе данных (на условном примере)

A – от реальности на основе взглядов исследователя, формирование представлений об объекте и предмете исследования, выделение основных понятий и связывающих их закономерностей через отнесение к ценности;

B – концептуализация: формирование эмпирической и математической систем, формирование и операционализация понятий с учетом «взаимодействия» исследователя и респондента;

C – операционализация понятий (более подробно ее можно посмотреть в работах П. Лазарсфельда, который разработал соответствующую теорию, сформулированную им на математическом языке и названную латентно-структурным анализом)[6]:

D – определение измеряемых объектов (построение и корректировка выборки), непосредственная реализация процедуры измерения;

E – реализация метода анализа данных;

F – интерпретация результатов применения метода.

[6] Статистические методы анализа информации в социологических исследованиях. М.: Наука, 1979. С. 249 – 266.

Статистическая закономерность, интересующая нас в данном случае больше всего, кроме того, **является результатом сжатия информации.** Результаты сжатия чаще всего выражаются в виде определенных характеристик частотных (вероятностных) распределений значений рассматриваемых признаков. Так, совокупность из 1 000 значений какого-либо признака может быть сжата до одного числа – среднего арифметического значения. Множество из 2 000 значений двух признаков можно сжать до одного числа – коэффициента парной связи между этими признаками. Совокупность из 10 000 значений десяти признаков может быть сжата до девяти коэффициентов регрессионного уравнения, связывающего один из рассматриваемых признаков с девятью другими и т.д. Сжатие информации предполагает ее потерю, а потому необходимо отслеживать, правомерны ли допускаемые потери (в частности, надо решить, те ли признаки выбраны для характеристики интересующих нас процессов, верно ли определен тип шкалы, правильно ли определено смысловое содержание чисел, получающихся в результате реализации алгоритма анализа данных, какую информацию мы можем позволить потерять при сборе и анализе данных и т.д.).

Сжатие должно быть таково, чтобы исследователь мог охватить взглядом массив: например, мы не сможем разобраться в результатах типологизации на 100 классов, каждый из которых характеризуется десятью признаками – и будем сжимать информацию дальше. То же можно сказать о ситуации, когда мы выявили 200 латентных факторов. Это должно учитываться в моделях.

1.2. Специфика математико-статистических методов применительно к социологической информации

Применение математики в социологии опирается на то, что мы считаем возможным:

а) выделить некоторый фрагмент реальности;

б) построить (посредством измерения) его математическую модель (т. е. получить исходные данные);

в) изучить эту модель традиционными для математики способами (применить тот или иной алгоритм анализа данных) и прийти к некоторым выводам (в результате анализа данных получить математический результат: точное значение коэффициента корреляции, параметры уравнения регрессии и т.д.);

г) проинтерпретировать эти выводы и получить таким образом новое знание.

Первые два этапа обычно относят к области измерения (шкалирования), последние два – к области анализа данных. Но все этапы тесно связаны друг с другом.

Выделенный фрагмент реальности называется эмпирической системой (ЭС). ЭС – это совокупность интересующих нас объектов, включая систему связывающих их отношений.

Процесс перевода всех компонент фрагмента реальности на формальный, математический язык, т. е. процесс измерения, позволяет нам перейти от ЭС к МС – математической системе (в социологии она может быть числовой или нечисловой).

Таким образом, **использование математических методов в процессе проведения социологического исследования позволяет достичь следующих целей:**

1. Побуждает исследователя четко формулировать свои представления об изучаемом объекте. Необходимым условием успешности здесь является комплексность анализа (использование группы методов). Так, желая сравнить величину связи между какими-либо признаками для разных совокупностей респондентов, мы, пытаясь построить математический критерий такой связи, вынуждены конкретизировать свои представления о ней. Это можно сделать многими способами (только коэффициентов парной связи между номинальными признаками известно более сотни; имея перед собой множество таких коэффициентов, мы можем понять, что есть наша связь в реальности)[7].

2. Позволяет абстрагироваться от большого количества реальных свойств изучаемых объектов.

3. Дает возможность получить содержательные выводы за счет расширения круга логических умозаключений.

---

[7] Типология и классификация в социологических исследованиях. М.: Наука. 1982. Гл. 7.

4. Дает возможность выявить скрытые механизмы взаимодействий при анализе огромных массивов информации (с которыми обычно и имеет дело социолог) и учете огромного количества факторов (определяющих любое общественное явление).

Типичной задачей, решаемой исследователем в процессе анализа анкетных массивов, является нахождение сочетаний значений признаков, которые детерминируют некоторое поведение респондента (скажем, голосование или неголосование на выборах). Результатом решения подобной задачи может служить, например, вывод, что среди мужчин старше 40 лет с высшим экономическим образованием, живущих в сельской местности, 95% проголосовало за лидера, т.е. что для респондентов, обладающих названными свойствами, характерна данная модель поведения. Но подобный вывод некорректен, т. к. мы не обнаруживаем всех требующихся групп респондентов. Здесь могут помочь специфические алгоритмы (например, алгоритмы типа AID, рассматриваемые ниже).

Таким образом, без применения математического аппарата трудно обойтись при решении практически любой социологической задачи.

1.3. Задачи математики применительно к социологической информации

Можно выделить две базовые **задачи, которые социолог ставит перед математической статистикой:**

– сжатие собранной эмпирической информации, направленное на вычленение скрытых в ней статистических закономерностей;

– решение проблемы соотнесения выборки и генеральной совокупности и построения репрезентативной выборочной совокупности. При изучении статистических закономерностей социолога всегда интересует задача перенесения полученных им результатов с той совокупности объектов, которая непосредственно была обследована (с выборки) на более широкую совокупность (генеральную).

Основными объектами изучения для математической статистики являются т. н. случайные величины. Это – функции, опре-

деленные на некоторых случайных событиях и принимающие числовые значения. В качестве типичного для социолога случайного события является выбор респондента. Случайными величинами могут служить признаки, определенные для этих респондентов. Выберем такой признак, как возраст: разные значения возраста (18, 24, 36, ... лет) это – разные значения нашей случайной величины. Случайная величина может быть и многомерной, когда ей отвечает несколько признаков, а ее значениями являются сочетания чисел – значений рассматриваемых признаков. Скажем, если наряду с возрастом мы будем учитывать пол (0 – мужчина, 1 – женщина) и зарплату (в рублях), то в качестве значений нашей трехмерной случайной величины могут выступать тройки чисел: (18, 0, 524), (36, 1, 1 200). При этом для каждой совокупности должна быть определена вероятность того, что, обследуя респондентов, социолог встретит значение из этой совокупности. **Вероятностью события** называют некоторую числовую характеристику степени возможности его появления в определенных, могущих повторяться неограниченное число раз, условиях.

Совокупность вероятностей встречаемости значений рассматриваемой случайной величины называется отвечающим ей **распределением вероятностей**, или просто ее распределением. Функция, задающая для определенных наборов значений случайной величины отвечающую им вероятность, называется **функцией распределения** случайной величины. На практике часто используется т.н. **функция плотности вероятности,** определяющая вероятность встречаемости каждого значения случайной величины (нормальное распределение, имеющее вид «колокола»).

Саму вероятность исследователь никогда не наблюдает и не может измерить. Это продукт нашего мышления, абстракция. Вероятность присуща генеральной совокупности, понятие которой также является абстракцией. Вместо вероятности исследователь обычно имеет дело с ее выборочной оценкой – относительной частотой встречаемости события. Чтобы было возможно использование аппарата математической статистики, необходимо частотные выборочные распределения расценивать как выборочные представления генеральных распределений вероятностей. Каждое такое распределение ассоциируется со случайной величиной.

Например, для выборки из 10 респондентов выборочное частотное распределение, отвечающее случайной величине «Удовлетворенность трудом», будет иметь вид, представленный табл. 1. С помощью тех же данных можно рассчитать и двухмерные распределения, одно из которых приведено в табл. 2 (для пары признаков).

Таблица 1

Выборочное представление частотного распределения случайной величины «Удовлетворенность трудом»

Значение признака	1	2	3	4	5
Частота встречаемости значения, %	30	30	10	10	20
Выборочная оценка вероятности P встречаемости значения	0,3	0,3	0,1	0,1	0,2

Таблица 2

Выборочное представление частотного распределения двухмерной случайной величины («Пол», «Удовлетворенность трудом»)

Пол	Удовлетворенность трудом					Итого
	1	2	3	4	5	
1	3	1	0	1	1	6
2	0	2	1	0	1	4
Итого	3	3	1	1	2	10

Математическая статистика позволяет выявить широкий круг статистических закономерностей (наборов параметров вероятностных распределений одномерных и многомерных случайных величин): меры средней тенденции, разброса значений случайных величин, связи между признаками и т.д. Результат, скажем, регрессионного анализа можно рассматривать как совокупность коэффициентов регрессии, которые в конечном итоге тоже являются некоторыми параметрами исходного многомерного распреде-

ления и т.д. Выборочные оценки параметров, рассчитанные на основе частотных распределений, называются **статистиками.**

Перейти от статистик к закономерностям генеральной совокупности можно, используя методы математического характера. Основные методы математической статистики обычно делят на две группы:

– методы статистической оценки параметров (способы расчета выборочных значений параметров и перехода от выборочных значений к генеральным; математическая статистика говорит о том, какими качествами эти оценки должны обладать, чтобы как можно более походить на их генеральные прообразы, и каким образом надо строить «хорошие» статистики, отражающие параметры вероятностных распределений);

– методы проверки статистических гипотез (оценка степени правдоподобности гипотезы о наличии некоторых соотношений между случайными величинами в генеральной совокупности на основании расчета определенных характеристик соответствующих выборочных распределений).

Правила переноса результатов с выборки на генеральную совокупность базируются на рассмотрении некоторых выборочных статистик как случайных величин и изучении определенных параметров их вероятности.

1.4. Сложности использования математических методов в социологии

Специалисты выделяют **ряд трудностей использования методов математической статистики в социологических исследованиях.** Их можно свести к следующим[8]:

I. Проблемы соотношения выборки и генеральной совокупности.

1. На практике нередко нарушаются условия вероятности совершения ожидаемого события.

Вероятность какого-либо события – это некая числовая характеристика степени возможности его появления в определен-

[8] Толстова Ю.Н. Анализ социологических данных. С. 38.

ных, могущих повторяться неограниченное число раз, условиях. Понятие вероятности имеет смысл, если рассматривается «круг явлений, когда при многократном осуществлении комплекса условий S доля той части случаев, когда событие А происходит, лишь изредка уклоняется сколько-нибудь значительно от некоторой средней цифры, которая, таким образом, может служить характерным показателем массовой операции (многократного повторения комплекса S) по отношению к событию А. Для указанных явлений возможно не только констатирование случайности события А, но и количественная оценка возможности его появления: вероятность того, что при осуществлении комплекса условий S произойдет событие А, равна р»[9].

В социологии само определение вероятности в некоторых ситуациях может стать бессмысленным: неясно, каков тот комплекс условий, повторение которого требуется, и будет ли он повторен вообще. Если в одной ситуации некое событие произошло, а в другой – нет, то мы практически никогда не узнаем, является это проявлением того, что вероятность данного события меньше единицы (реализовав много ситуаций и подсчитав долю тех, в которых наше событие свершилось, мы тем самым получим оценку соответствующей вероятности), либо мы имеем дело со следствием того, что разные ситуации отвечают разным комплексам условий, задающих вероятность, и что поэтому вероятности нашего события в этих ситуациях различны.

2. Не всегда ясно, какова изучаемая генеральная совокупность. Социолог имеет в своем распоряжении всего одну выборку, не всегда корректно рассчитанную. Методы поиска закономерностей «в среднем» в подобной ситуации нельзя отнести к математическим в полной мере. Социологи все же для удобства предполагают, что гипотетическая генеральная совокупность существует и что имеющиеся в нашем распоряжении выборочные частоты – это хорошие оценки соответствующих генеральных вероятностей, а потому работает с этим распределением так, как правила математической статистики предписывают работать с распределением вероятностей.

[9] Гнеденко Б.В. Курс теории вероятностей. М.: Наука, 1965. С. 15.

3. Для многих методов отсутствуют разработанные способы перенесения результатов их применения с выборки на генеральную совокупность (о чем скажем ниже).

4. Методы переноса результатов с выборки на генеральную совокупность обычно базируются на серьезных теориях. Если такой теории нет, социолог или интуитивно выбирает генеральную совокупность, или использует ЭВМ для создания распределений искусственным путем (такой подход – Bootstrap – активно развивается на Западе).

5. Перенос результатов с выборки на генеральную совокупность может быть затруднен из-за «ремонта» выборки. Тут тоже может помочь моделирование данных на ЭВМ.

II. Отсутствие строгих обоснований возможности применения конкретных методов математической статистики.

Для некоторых методов, показавших свою эффективность при решении практических задач, отсутствуют строгие доказательства корректности их использования. Например, для применения метода регрессионного анализа к данным, полученным в результате дихотомизации номинальных признаков. Но методы используются, несмотря на их некорректность. И для обозначения совокупности таких некорректных методов, для отделения их от строгих математико-статистических подходов, был введен термин «анализ данных». Поэтому, заметим, особое значение приобретает проблема обоснованности получаемых с их помощью выводов.

III. Использование шкал низких типов.

Интересующие социолога данные, как правило, получены по шкалам низких типов[10]. Шкалами низкого типа считают шкалы номинальные и порядковые, а шкалами высокого типа – интервальные и шкалы отношений. Шкалы низкого типа (и получаемые с их помощью данные) называют также качественными, а

[10] Толстова Ю.Н. Анализ социологических данных. С. 40 – 48.

шкалы высокого типа (и соответствующие данные) – количественными, или числовыми[11].

Номинальной шкалой мы называем такую шкалу, с помощью которой стремимся отразить в числах только некоторое отношение равенства-неравенства между изучаемыми объектами. Типичным признаком, значения которого обычно получаются именно по номинальной шкале, является профессия респондента. Если одному респонденту приписано значение «3» («токарь»), а другому – значение «4» («пекарь»), то, имея в руках эти числа, мы можем быть уверенными в том, что рассматриваемые объекты в интересующем нас отношении различны (респонденты имеют разные профессии), но больше ничего мы о них сказать не можем.

При использовании *порядковой* шкалы мы ставим целью отобразить не только некоторое отношение равенства-неравенства между реальными объектами, но и содержательное отношение порядка между ними. В качестве примера может служить анкета с вопросом «Удовлетворены ли Вы Вашей работой (ходом реформ, президентом РФ…)?» и веером из 5 (3, 7 и т.д.) вариантов ответов от «Совершенно не удовлетворен» до «Вполне удовлетворен», которым ставятся в соответствие числа от 1 до 5 (от 1 до 3, от 1 до 7, от –3 до +3 и т.д.). Здесь мы при осуществлении шкалирования ставим целью отобразить в числах не только отношение равенства респондентов по их удовлетворенности объектом, но и отношение порядка между респондентами по степени «накала» их эмоций, направленных в адрес этого объекта. И если окажется, что одному респонденту приписано число «2», а другому – «4», то мы будем полагать, что упомянутый «накал» второго респондента не просто не равен «накалу» первого, но больше такового[12].

Для чисел, полученных по шкалам низких типов, не имеет смысла большинство традиционных операций с числами. Так, вряд ли найдется человек, усматривающий что-то рациональное в утверждениях: «Среднее арифметическое значение профессий для рассматриваемой совокупности респондентов равно 3,2, и оно меньше аналогичного среднего значения для другой совокупности, равного 3,9». Что значит величина 3,2? То, что некий

[11] Толстова Ю.Н. Указ. соч.

[12] Там же.

средний, наиболее типичный респондент на 20% является токарем, а на 80% – пекарем[13]?

В *интервальных* шкалах полученные данные похожи на действительные числа, но все же таковыми не являются. Они отображают в числовых отношениях не только некоторые эмпирические отношения равенства и порядка, но и структуру эмпирических интервалов – отношения равенства и порядка для расстояний между объектами.

Возможности использования математической статистики для изучения данных, полученных по шкалам низких типов, подробнее изучаются статистикой объектов нечисловой природы[14].

IV. Необходимость соотнесения модели метода с содержанием социологической задачи.

Если для решения социологической задачи существует некоторый математический метод, то этот метод практически никогда не бывает единственным. Например, существует много мер средней тенденции, разброса частотного распределения значений признака. Для измерения связи даже между двумя номинальными признаками могут служить более 100 коэффициентов. Еще большее разнообразие присуще сложным методам изучения многомерных распределений (SPSS в одном алгоритме классификации CLUSTER предусматривает использование 6 способов измерения расстояний между объектами и 7 способов расстояний между классами, т. е. 42 варианта классификации). И за каждым методом – свое понимание изучаемого явления (средней тенденции, разброса, связи и т.д.).

Приведем пример расчета мер средней тенденции, чтобы показать, что такой выбор может диктовать нам содержание задачи (приведен Ю.Н. Толстовой): «Опишем некоторую задачу о моде в житейском смысле этого слова. Предположим, что модельер должен определить, какая длина должна быть у очередной модели женских юбок, выпускаемых фабрикой, и для этой цели опра-

[13] Толстова Ю.Н. Указ. соч.

[14] Подробнее об этом см.: Орлов А.И. Общий взгляд на статистику объектов нечисловой природы // Анализ нечисловой информации в социологических исследованиях. М.: Наука: 1985. С. 58 – 92.

шивает женщин рассматриваемого региона, просит их указать "любимую" длину. Если мы в качестве длины, рекомендуемой фабрике, укажем медиану соответствующего распределения, то тем самым окажемся перед риском выпустить неходовой товар: половина женщин решит, что юбка для них слишком коротка, а половина – что чересчур длинна. Покупать продукцию фабрики никто не захочет. А вот если в качестве меры средней тенденции мы используем моду, то удовлетворим женщин, выразивших наиболее часто встречающееся мнение»[15].

Терстоун, предлагая свой метод построения шкалы для измерения установки, рекомендовал на последнем этапе процедуры, при расчете приписываемого каждому респонденту итогового балла, использовать медиану в качестве среднего значения весов тех суждений, с которыми этот респондент согласился (а не среднее арифметическое).

Дэйвисон рассматривает задачу изучения пространства восприятия респондентами некоторых объектов с помощью многомерного шкалирования. Предлагается способ построения матрицы близости между объектами на основе своеобразного опроса респондентов, и для усреднения соответствующих мнений рекомендуется использовать среднее геометрическое.

Тема 2
Общая характеристика процедуры анализа данных

2.1. Социологические данные

Под **данными** мы будем понимать первичную информацию, полученную в результате социологического исследования: ответы респондентов, оценки экспертов, результаты наблюдения и т.п., совокупность значений переменных, приписанных единицам исследования – объектам.

[15] Толстова Ю.Н. Указ. соч. С. 48.

Социологические данные это[16]:

– совокупности чисел, характеризующих объекты исследования – производственные характеристики предприятий, возраст людей, оценки выпускниками престижности профессий и т.д.;

– индикаторы определенных отношений между рассматриваемыми объектами (например, симпатия-антипатия в малой группе);

– результаты попарных сравнений респондентами каких-либо объектов;

– совокупности определенных высказываний (оценки политики правительства; письма читателей газеты в редакцию; фрагменты из журнальных статей и т.д.);

– тексты документов;

– зафиксированные результаты наблюдения за невербальным поведением людей и т.п.

С социологическими данными можно производить следующие операции:

а) подготавливать их для обработки, шифровать, кодировать и т.д.;

б) обрабатывать (вручную или с помощью компьютера): табулировать, рассчитывать многомерные распределения признаков, классифицировать и т.д.;

в) анализировать;

г) интерпретировать[17].

Наиболее часто в социологических исследованиях данные представляют собой совокупность значений признаков (характеристик, переменных, величин) объекта.

Признак – некоторое общее для всех объектов качество, конкретные проявления которого (значения признака; их называют также альтернативами, градациями) могут меняться от объекта к объекту (пол, возраст респондентов, их удовлетворенность своим трудом). В качестве значений признака «возраст» могут выступать 25 лет, 48 лет, 21 год. Признаки – наши абстрактные

---

[16]Толстова Ю.Н. Указ. соч. С. 3 – 5.

[17] Методическое пособие социолога-практика / под ред. Д.А. Шевченко, А.И. Кравченко. Советская социологическая ассоциация АН СССР. М., 1990. С. 86.

идеальные конструкции. В общественных науках соответствующий процесс абстрагирования является иногда очень непростым. Основными этапами абстрагирования являются выделение понятий и осуществление их операционализации. На практике проблему операционализации чаще всего разделяют на:

– выбор признаков, являющихся индикаторами понятий;

– выбор набора значений каждого признака (выбрав в качестве одного из индикаторов признак «возраст», мы можем считать его «непрерывным» и просить каждого респондента указывать целое число прожитых лет; можем приписывать респонденту число от 1 до 5 в зависимости от того, в какой возрастной интервал респондент попадает: от 15 до 25 лет, от 25 до 35 лет, … , старше 55 лет; разделим всех людей на две группы – до 30 лет и старше и т.д.).

Социолог рассматривает ситуацию, когда каждый изучаемый объект предстает перед ним в виде последовательности чисел – значений признаков. Такие данные обычно задаются в виде таблицы (матрицы) «объект-признак», строки которой отвечают объектам (например, респондентам), а столбцы – признакам (например, каждый столбец – это ответы респондентов на один из вопросов анкеты): см. табл. 3.

При использовании методов многомерного анализа данных ту же информацию об исходных объектах представляют в виде фрагмента т.н. признакового пространства: осям такого пространства отвечают рассматриваемые признаки, а каждый объект представлен в виде точки, координатами которой служат значения для этого объекта признаков, отвечающих осям. Пример двухмерного признакового пространства, оси которого отвечают признакам «Возраст» и «Удовлетворенность трудом», а координаты объектов – данные табл. 3, приведен на рис. 2.

Обработкой социологической информации называют математико-статистическое преобразование данных, которое делает их компактными, пригодными для анализа и интерпретации.

Таблица 3

Матрица «объект-признак»

Номер объекта (респондента)	Наименование признака		
	Пол (0 – муж., 1 – жен.)	Возраст, лет	Удовлетворенность трудом (1 – совершенно не удовлетворен, …, 5 – полностью удовлетворен)
1	0	25	1
2	0	31	2
3	0	18	5
4	1	24	2
5	0	18	1
6	0	38	4
7	1	41	3
8	1	50	1
9	1	54	2
10	1	19	5

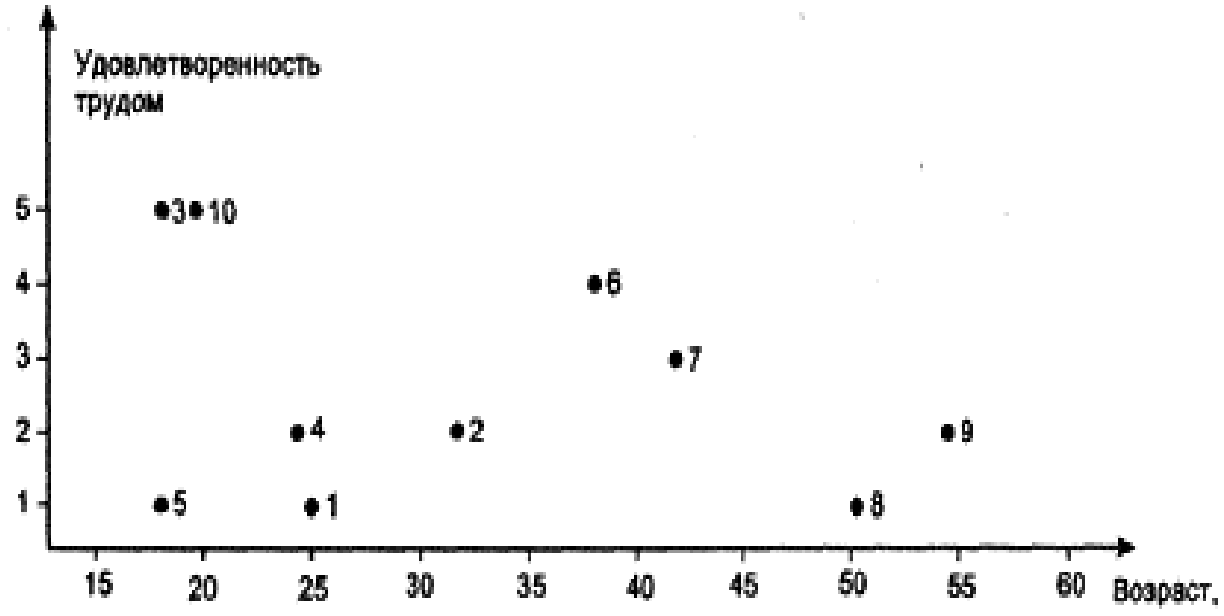


Рис. 2. Двухмерное признаковое пространство «возраст – удовлетворенность трудом»

2.2. Общие принципы анализа данных

По мнению известного российского социолога В.А. Ядова, «анализ собранной информации – самый увлекательный этап исследования»[18]. Этому этапу исследования посвящен огромный пласт специальной литературы[19].

Ю.Н. Толстова указывает на существование четырех различных, связанных между собою смыслов понятия **анализ данных** в социологии:

а) совокупность действий, совершаемых в процессе изучения полученных эмпирических данных для того, чтобы сформировать представление о характеристиках изучаемого явления;

б) процесс изучения статистических данных с помощью неких приемов, математических методов и моделей с целью более удобного и наглядного их представления, что позволяет наиболее обоснованно интерпретировать изучаемое явление;

в) понятие, тождественное прикладной статистике;

г) процедуры «свертывания» информации, которые не допускают формального алгоритмического подхода[20].

Основная **цель анализа данных –** выявление (подтверждение, корректировка) каких-то интересующих исследователя статистических закономерностей или сжатие, усреднение содержащейся в данных информации.

В задачу поиска закономерности включают:

– *объяснения* интересующего исследователя явления (смысл объяснения состоит в подведении объясняемого явления под какой-либо закон);

– *описание* исходных данных для того, чтобы исследователь мог сориентироваться в большом объеме данных, понять, какие закономерности скрываются за интересующими его данными, ка-

---

[18] Ядов В.А. Социологическое исследование: методология, программа, методы. Самара: Самарский университет, 1995. С. 202.

[19] В.А. Ядов приводит список, состоящий из 314 наименований отечественной и зарубежной литературы плюс аннотированный список из более 70 названий (см.: Ядов В.А. Социологическое исследование: методология, программа, методы. С. 275 – 285, 309 – 329).

[20] Социологическое исследование: методы, математика и статистика // Социология: Словарь-справочник. М., 1991. Т. 4. С. 7 – 9.

кими признаками эти закономерности должны описываться, возможно ли подобрать соответствующие признаки и т.д. Описание обычно достигается с помощью самых простых способов сжатия исходных данных. Примеры: доля женщин в изучаемой совокупности; средний возраст респондентов; величина разброса респондентов по возрасту (например, выраженная в виде соответствующей дисперсии); наиболее часто встречающаяся среди респондентов профессия; нижний уровень дохода 10 % самых богатых респондентов и т.д. Совокупность наиболее употребительных приемов получения закономерностей, описывающих изучаемое множество объектов, называется **описательной, или дескриптивной, статистикой;**

– *предсказание* на основе выявленной закономерности того или иного явления с помощью сложных алгоритмов (алгоритмы регрессионного анализа);

– *понимание* изучаемого явления. Оно обычно достигается с помощью мягких методов исследования [21].

Общие **принципы** анализа социологической информации можно свести к следующим: упорядочение, уплотнение, компактное описание собранной информации. Они реализуются в ходе **аналитических процедур.**

Собранная в ходе полевого этапа первичная социологическая информация не структурирована, а потому не поддается непосредственному изучению. Упорядочение информации осуществляется с помощью статистической группировки данных и типологизации информации[22].

Метод *статистической группировки* заключается в том, что обследуемая совокупность расчленяется на однородные группы (отдельные единицы которых обладают общим для всех признаком).

При **группировке по количественным признакам** (возраст, стаж работы, размер дохода) весь диапазон изменения переменной разбивают на определенные интервалы с последующим подсчетом числа единиц, входящих в каждый из них.

---

[21] Толстова Ю.Н.. Анализ социологических данных. С. 34.

[22] Ядов В.А. Социологическое исследование: методология, программа, методы. С. 202 – 207.

При **группировке по качественным признакам** каждая из единиц анализа относится к одной из выделенных градаций с тем, чтобы суммарное число единиц анализа, отнесенных ко всем градациям, было бы равно общей численности изучаемой совокупности.

Метод *типологизации информации* представляет собой обобщение признаков социальных явлений на основе идеальной теоретической модели и по теоретически обоснованным критериям[23]. В качестве примера можно привести исследование политической ориентации жителей Ярославской области, в ходе которого выделяются такие типы политической ориентации, как демократы, либералы, коммунисты, националисты и т.п.

Математический аппарат, используемый в эмпирической и прикладной социологии, предлагает для выявления связи между явлениями, определения ее направления и силы большое число специализированных процедур. Выбор их для конкретного исследования зависит от задач исследования, от уровня подготовки исследователя, от корректности целей.

Тема 3
Анализ одномерных распределений

3.1. Необходимость анализа одномерных распределений в социологии

Основной объект изучения математической статистики – случайная величина – превращается в привычный социологу признак (пол, возраст, удовлетворенность жизнью). В качестве случайных событий рассматриваются только те, которые состоят в том, что какие-то признаки принимают определенные значения (например, событие может состоять в следующем: взяв анкету,

[23] Ядов В.А. Социологическое исследование: методология, программа, методы. С. 208.

исследователь увидел, что ему «попался» мужчина старше 30 лет, крайне недовольный жизнью). В качестве оценки вероятности того или иного события выступает относительная частота его встречаемости в конкретной изучаемой социологом выборке (событие имеет вероятность 0,25, если доля мужчин с указанными свойствами в изучаемой выборке составляет 25%).

Социолог практически всегда начинает свою работу с некоторого описания интересующей его совокупности объектов. Для этой цели чаще всего используется расчет частотных распределений (одномерных, двухмерных, многомерных), разных показателей среднего уровня значений какого-либо признака, а также индикаторов разброса таких значений.

Вначале необходимо описать данные по каждой из переменных (*описательная статистика).* Соответствующие таблицы называют *линейными* или *одномерными* распределениями. Мы можем анализировать частотное распределение значений рассматриваемого признака, т. е. выборочное представление изучаемой одномерной случайной величины. Такие описания позволяют исследователю лучше сориентироваться в проблематике, скорректировать перечень проверяемых гипотез, уточнить представления об объекте и предмете исследования. Описательные статистические данные – это данные, полученные в результате математического суммирования многочисленных наблюдений[24].

Обычно для обобщенного описания используют два основных типа анализа:

а) измерение *центральной тенденции* (наиболее часто встречаемых значений переменных в линейных распределениях);

б) измерение разброса, или *дисперсии* (плотность или слабость распределения значений переменной вокруг наиболее общего, среднего или центрального значения).

Однако при выборе типа анализа мы должны принимать во внимание *шкалу,* с помощью которой производилось измерение переменной (см. Тема 4).

[24] Добреньков В.И., Кравченко А.И.. Методы социологического исследования. М.: Инфра-М, 2006. С. 193.

3.2. Меры средней тенденции

Для одномерных случайных величин можно вычислить меры средней тенденции: в социологии наиболее часто используются математическое ожидание, мода и квантили (наиболее употребительным квантилем является медиана). Они являются параметрами распределения вероятностей.

Выборочные оценки параметров распределения делятся на точечные, когда для выборочных данных находится одно значение, служащее оценкой генерального параметра, и интервальные, когда на базе выборочной точечной оценки параметра строится так называемый доверительный интервал. Покажем выборочные точечные оценки указанных параметров. Определенная на выборке переменная, значениями которой служат точечные оценки какого-либо параметра, называется *статистикой, отвечающей этому параметру*[25].

Пусть x_1, x_2, ..., x_N – выборочные значения рассматриваемого признака (N – объем выборки). Статистикой, отвечающей математическому ожиданию, является *среднее арифметическое* значение признака (значение наиболее типичного для группы человека):

$$\overline{x} = \frac{(x_1 + x_2 + ... + x_N)}{N}.$$

Квантиль – это такое значение признака q, которое делит диапазон его изменения на две части так, чтобы отношение числа элементов выборки, имеющих значение признака, меньшее q, к числу элементов, имеющих значение признака, большее q, было равно заранее заданной величине. Наиболее популярными квантилями являются: *квартили*, разбивающие диапазон изменения признака на 4 равнонаполненные части; *децили* – на 10 равнонаполненных частей; *процентили* – на 100 частей (рис.3).

[25] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976. С. 114.

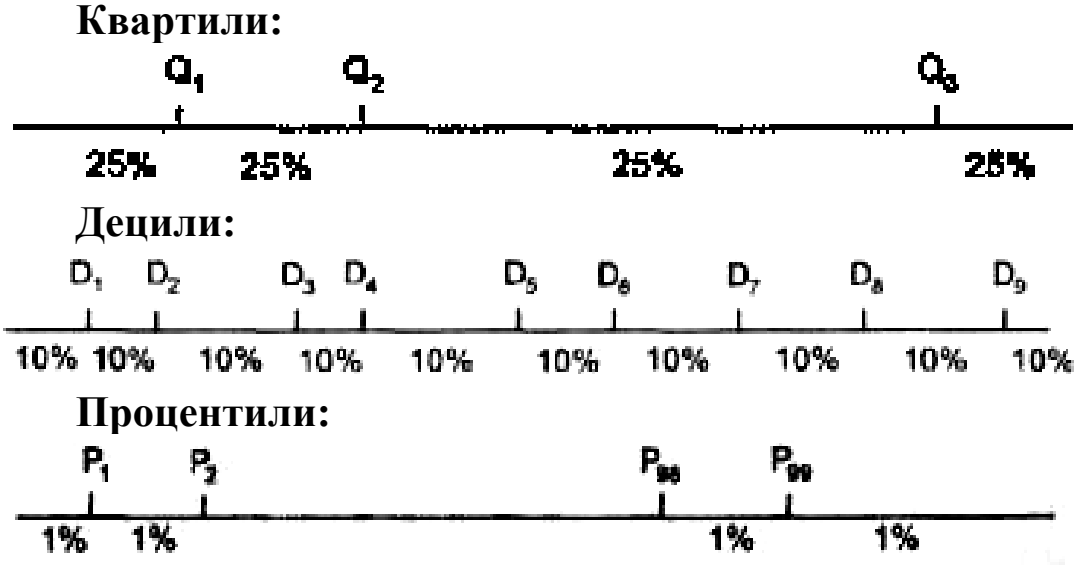


Рис. 3. Наиболее употребляемые квантили

Величина процента, указанная под интервалом, означает долю объектов выборки, попавших в этот интервал. Квантилями социолог пользуется очень часто. Например, 10% наиболее богатых людей имеют месячный доход свыше 100 000 рублей, а 10% наиболее бедных – ниже 1 000 рублей. 100 000 рублей – это девятый дециль D_9, а 300 рублей – это первый дециль D_1.

Выборочная медиана – это значение рассматриваемого признака, которое делит отвечающий этому признаку *вариационный ряд* (т. е. последовательность значений признака, расположенных в порядке их возрастания) пополам: половина всех выборочных значений признака меньше нее, а половина – больше. Допустим, есть 2 группы, в одной из которых медиана признака «доход» равна 500 рублей, а в другой – 5 000 рублей. Ясно, что вторая группа «в среднем» гораздо богаче первой. Обычно, построив вариационный ряд, полагают, что при нечетном числе элементов в выборке медиана равна центральному члену ряда, а при четном – точке, отвечающей середине расстояния между двумя центральными членами[26].

$$Me = Q_2 = D_5 = P_{50}.$$

[26] Толстова Ю.Н. Анализ социологических данных. С. 76 – 78.

Вычисление медианы имеет смысл только для порядкового и интервального признаков. В случае же, когда медиана вычисляется как середина между двумя шкальными значениями, мы делаем еще одно предположение – о том, что наш порядковый признак может принимать значения, лежащие между используемыми пунктами шкалы.

Можно рассчитывать медиану и с помощью построения кумуляты. Это также опирается на предположение о непрерывности рассматриваемого признака или о том, что внутри каждого интервала значения признака распределены равномерно (рис. 4.).

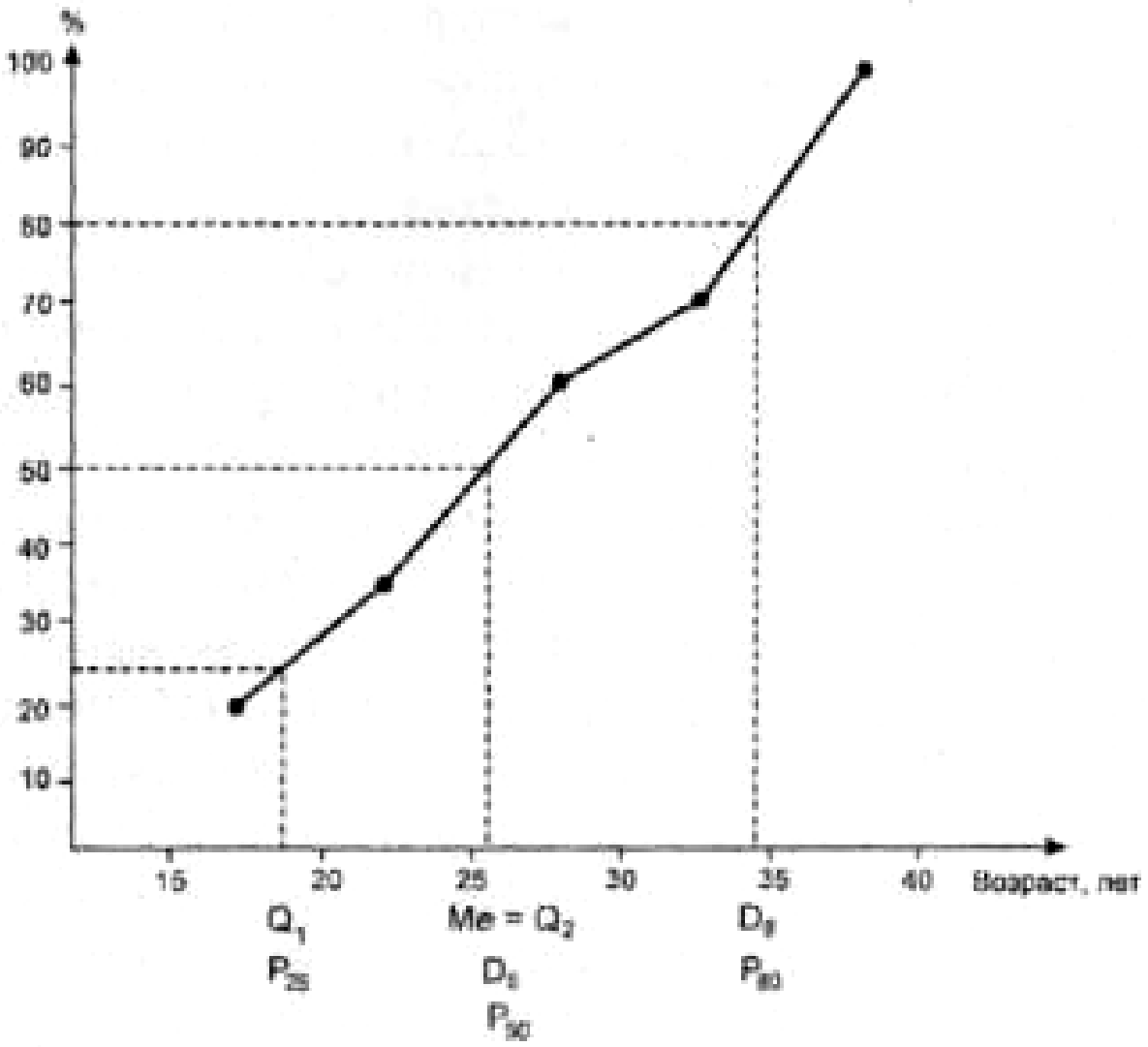


Рис. 4. Один из возможных способов расчета квантилей

Опишем разные способы расчета медианы на примере.

Предположим, что для 10 школьников значения коэффициента IQ, определенные с помощью шкалы интеллекта Стенфорда-Бине, оказались следующими:

113, 120, 119, 115, 122, 126, 120, 112, 120, 119.

Прежде всего необходимо определить тип используемой шкалы. Будем считать шкалу – интервальной с интервалами (128 – 127); (113 – 112) и т. п.

Известно, что значением коэффициента может быть любое целое число от 0 до 150. Покажем, какими способами можно рассчитать медиану этого распределения.

А. Выборка – это и есть генеральная совокупность. Тогда медиану целесообразно найти с помощью вариационного ряда:

112, 113, 115, 119, 119, 120, 120, 120, 122, 126.

Me = 119,5

В таком случае естественной будет функция распределения, изображенная на рис. 5.

Б. Рассмотрим другую функцию распределения, в основе которой лежат два предположения:

– в качестве значения переменной служит любое действительное число из рассматриваемого диапазона. После принятия указанного предположения функцию распределения естественно представлять в виде отрезков построенной ломаной линии, соединяющих левые концы стрелок (рис. 6);

– объекты в каждом заданном выборкой интервале накапливаются равномерно. Так, если в процессе построения графика накопленных частот (выборочного аналога функции распределения) в точке горизонтальной оси 115 у нас 30% объектов, а в точке 119 – уже 50%, то мы считаем, что 20% объектов, попавших в интервал (115, 119), равномерно распределены в этом интервале и соответствующий фрагмент функции распределения есть отрезок прямой, соединяющий точки (115, 30) и (119, 50). Медиана в таком случае находится традиционным способом, отраженном на рисунке (она равна 119, а не 119,5).

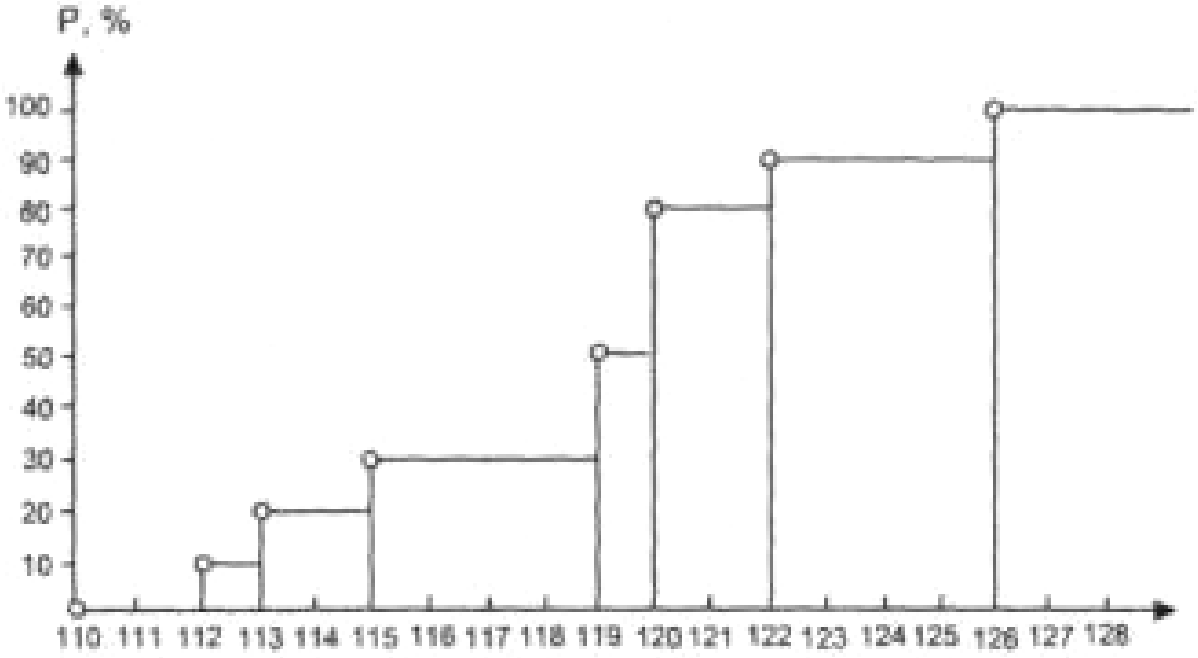


Рис. 5. Вид функции распределения при отождествлении выборки с генеральной совокупностью

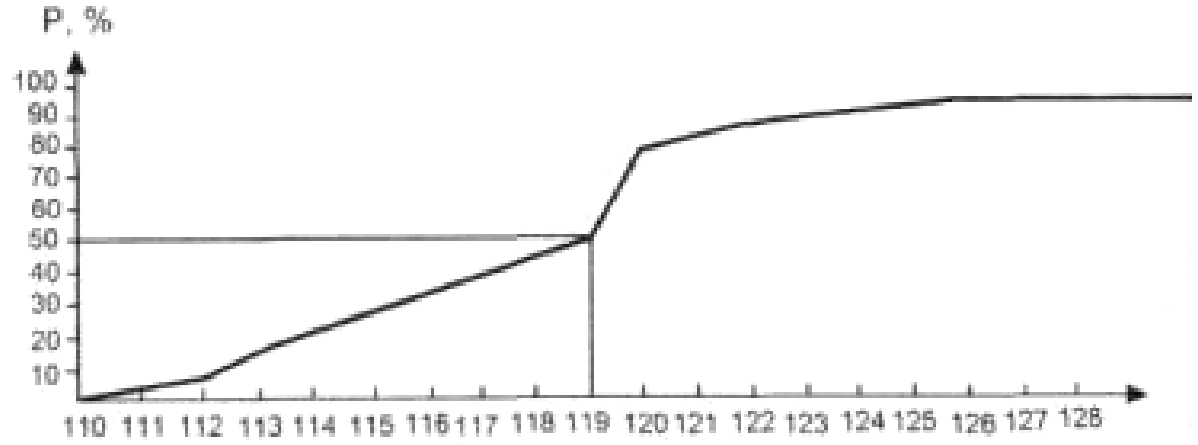


Рис. 6. Вид функции распределения при предположениях о непрерывности рассматриваемой случайной величины и равномерном накоплении единиц совокупности в каждом заданном выборкой интервале

Социолог обычно разбивает диапазон изменения рассматриваемого признака на интервалы и полагает, что в действительности для него при рассмотрении конкретного объекта имеет смысл не то, какое именно значение признака этому объекту отвечает, а то, в какой интервал это значение попадает. При построении вы-

борочного представления функции распределения доля объектов, отвечающих какому-либо интервалу, откладывается от любой точки последнего. Вид функции распределения при предположениях о непрерывности рассматриваемой случайной величины, заданном разбиении на интервалы диапазона ее изменения, отнесении точки стыка двух интервалов направо, равномерном накоплении единиц совокупности в промежутке от середины одного интервала до середины другого представлен на рис. 7. В данном случае медиана будет равна 117,5.

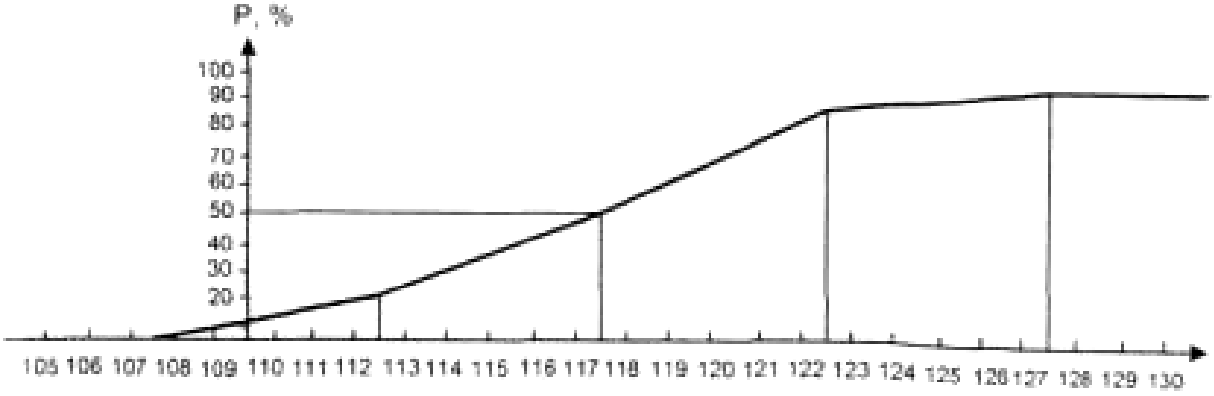


Рис. 7. Вид функции распределения для медианы Me =117,5

Вид функции распределения при предположениях о непрерывности рассматриваемой случайной величины, заданном разбиении на интервалы диапазона ее изменения, отнесении точки стыка двух интервалов направо, равномерном накоплении единиц совокупности в каждом интервале представлен на рис. 8. В данном случае медиана будет равна 119.

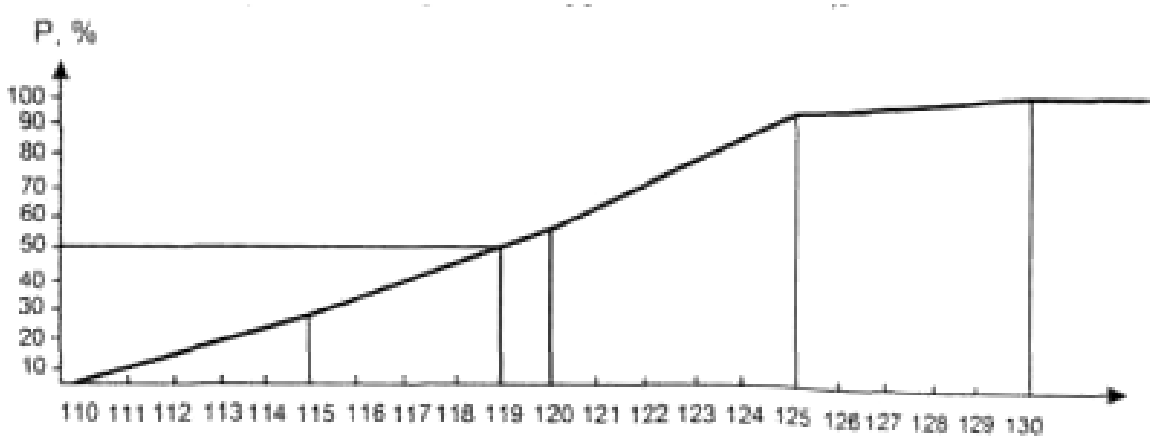


Рис. 8. Вид функции распределения для медианы Me = 119

Модой называется наиболее часто встречающееся значение признака. Ее можно рассчитывать для признаков, измеренных по шкалам любых рассматриваемых нами типов. Сравнивая, скажем, распределение по профессиям, рассчитанное для двух регионов – Ивановской и Тюменской области, мы можем прийти, например, к выводу, что в первой наиболее распространенная профессия – ткачиха, а во второй – нефтяник. Этот вывод означает, что ткачиха – модальное значение профессии для жителей Ивановской области, а нефтяник – для Тюменской[27].

Так мы находим линейные закономерности.

Любая статистическая закономерность – это своего рода *сжатие исходных данных*. Так, при использовании среднего арифметического мы вместо набора из 1 000 значений возрастов мы получили одно число – 32,4, средний возраст респондентов совокупности. Совокупность из тысячи чисел сжата в одно число. Указанное сжатие означает *потерю информации.*

Среднее арифметическое предполагает использование интервальной шкалы, т.к. это такое значение признака, для которого сумма расстояний от него до объектов, имеющих большее значение, равна сумме расстояний до объектов, имеющих меньшее значение:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_N - \bar{x}) = 0.$$

Для порядковых шкал большинство подобных соотношений не будут формально адекватными. Номинальные шкалы требуют особого пояснения. Использование среднего арифметического для чисел, например, кодов профессий респондента, является бессмысленным. Тем не менее бывают случаи, когда и для номинальных данных оказывается возможным использование этой статистики (дихотомические номинальные признаки, принимающие два значения – 0 и 1). Рассмотрим самый популярный дихотомический признак – пол респондента: 0 – мужчина, 1 – женщина.

Предположим, что у нас 10 респондентов со следующими значениями пола[28]:

---

[27] Толстова Ю.Н. Указ. соч. С. 76 – 78.

[28] Там же.

0, 0, 1, 1, 1, 0, 0, 0, 0, 1.

Среднее арифметическое в данном случае равно 0,4. Это не пол «среднего человека» (типичным представителем совокупности является человек, на 40% являющийся женщиной, на 60% мужчиной). Это означает, что в совокупности имеется 40% людей с единичным значением рассматриваемого признака (40% женщин). Так можно использовать числовой анализ для изучения номинальной информации.

3.3. Дисперсия

Используя для описания выборки только меру средней тенденции, исследователь рискует сильно ошибиться. Например, если изучаемый признак – возраст, то две совокупности людей из 6 человек каждая, характеризующиеся следующими значениями возраста:

10, 10, 10, 50, 50, 50
30, 30, 30, 30, 30, 30,

будут иметь одинаковое среднее арифметическое 30. Но это будут совсем разные совокупности, что подтверждает оценка степени разброса значений возраста в каждой: в первой разброс большой, во второй он отсутствует.

Самой известной мерой разброса количественного признака является его **дисперсия:**

$$\sigma^2 = \frac{((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2)}{N - 1}$$

(в знаменателе величина объема выборки уменьшается на единицу для того, чтобы сделать соответствующую точечную выборочную оценку дисперсии несмещенной). Эта статистика – только для **интервальных** шкал. Для **порядковых** шкал используют разницы между квантилями (квартильный размах $Q_3 - Q_1$). Для

номинальных признаков некорректным является использование всех приведенных выше мер разброса.

Чем ближе распределение к равномерному – тем разброс больше, чем дальше от равномерного – тем разброс меньше. Известны две меры разброса, опирающиеся на этот принцип – мера качественной вариации и энтропийный коэффициент разброса.

3.4. Мера качественной вариации

Предположим, что мы организовали танцевальный кружок из 10 человек и пытаемся путем перебора различных вариантов формирования разнополых пар найти такие, где мужчина и женщина наиболее удачно подходят друг другу как танцоры (табл. 4)[29].

Мы видим, что наибольшее количество пар можно организовать, когда распределение по полу равномерно (т.е. количество мужчин равно количеству женщин – когда разброс членов кружка по полу максимален).

Таблица 4

Зависимость количества пар из разнородных элементов от степени однородности распределения

Количество мужчин в кружке	Количество женщин в кружке	Количество возможных танцевальных пар
0	10	0
1	9	9
2	8	16
3	7	21
4	6	24
5	5	25
6	4	24
7	3	21
8	2	16
9	1	9
10	0	0

[29] Толстова Ю.Н. Указ. соч.

Уровень разброса респондентов по полу и в остальных случаях четко коррелирует с количеством пар из разнородных элементов: чем больше разброс, тем больше пар можно составить. В этой мере разброса – мере качественной вариации – «ядро» составляет величина, равная количеству упомянутых пар. Поясним на примере способ расчета этой меры (табл. 5).

Таблица 5

Расчет коэффициента качественной вариации

Наименование градации рассматриваемого номинального признака	A	B	C
Частота встречаемости градации	30	20	70

Вычислим коэффициент по следующей формуле:

$$J = \frac{30\times 20 + 30\times 70 + 20\times 70}{40\times 40 + 40\times 40 + 40\times 40}.$$

В числителе дроби стоит число, равное количеству пар, которые можно составить из разнокачественных элементов: произведение 30 и 20 – количество пар, первый элемент, который обладает свойством A, а второй – свойством B; 30 и 70 – то же для свойств A и C; 20 и 70 – для свойств B и C. Числитель отражает существо разброса, но не является его мерой. Границы его изменения зависят от объема выборки, от величины конкретных частот. Поэтому, ограничившись числителем, мы теряем возможность сравнивать меры разброса разных совокупностей: число, отвечающее большому разбросу в малой выборке, может говорить о несущественном разбросе в большой выборке. Это недопустимо, т.к. анализ данных связан со сравнением разных совокупностей объектов.

Если дополнить данные табл. 5 данными, уменьшенными в 10 раз, то получим две разные выборки, характеристики которых отражены в табл. 6.

При объеме выборки в 12 человек (и при трех градациях признака) максимальное количество пар из разнородных элементов равно 48 (перемножаем 4 и складываем произведения 4·4 + 4·4 + + 4·4). Для выборки в 12 человек число 48 говорит о максималь-

ном разбросе. А при объеме выборки в 120 человек (при тех же трех градациях) такого малого количества пар не может быть даже при самом минимальном (но ненулевом) разбросе. Такой минимальный разброс будет иметь место, если какое-то одно значение встречается 119 раз, а другое – всего один (при отсутствии третьего значения). Количество же пар из разнородных элементов в таком случае будет равно 119, что больше 48.

Таблица 6

Зависимость величины меры качественной вариации от объема выборки

Наименование градации рассматриваемого признака	Число респондентов (частота) в первой выборке – 120 чел.	Гипотетические частоты, отвечающие максимальному значению J	Число респондентов (частота) во второй выборке – 12 чел.	Гипотетические частоты, отвечающие максимальному значению J
A	30	40	3	4
B	20	40	2	4
C	70	40	7	4

Если мы будем пользоваться только числителем дроби, выражающей коэффициент J, то в одном случае число 48 говорит о максимальном разбросе, а в другом число 119 – об отсутствии разброса. Мы не можем сравнить коэффициенты разных совокупностей. Поэтому в числитель помещают формулу, выражающую суть строящегося коэффициента, а в знаменатель – максимально возможное значение этого коэффициента для рассматриваемой ситуации (она определяется объемом выборки и количеством градаций признака). Показатель рассматривается в интервале от 0 до 1 (иногда от –1 до +1, как в случае коэффициента корреляции). Такая процедура называется **нормировкой коэффициента** (деление числителя на аналогичную сумму произведений, отвечающую равномерному распределению, т.е. распределению, когда все градации признака встречаются с одинаковой частотой). Общая формула коэффициента J[30]:

[30] Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. Киев: Наукова Думка. 1982. С. 84.

$$J = \frac{2K}{N^2(k-1)} \cdot \sum_{i-1}^{k-1} \sum_{j-i+1}^{k} n_i n_j ,$$

где N – объем выборки; k – количество градаций рассматриваемого признака; n_i и n_j – соответственно частоты встречаемости i-й и j-й градаций.

Если мы имеем дело с дихотомическим признаком, принимающим два значения – 0 и 1, то, вычислив обычную дисперсию, мы получим коэффициент качественной вариации.

3.5. Энтропийный коэффициент разброса

Степень неопределенности распределения случайной величины Y определяется с помощью энтропии этого распределения. Пусть случайная величина Y принимает значения 1, 2, ..., k с вероятностями, равными P_1, P_2, ..., P_k (вероятность отождествляется с относительной частотой встречаемости этого значения). Введем обозначение:

$$P_j = P\ (Y = j).$$

Энтропией случайной величины Y (распределения) называется функция **(формула Больцмана)** вида

$$H(Y) = -\sum_{j-1}^{K} P_j \log P_j ,$$

где основание логарифма произвольно.

Пусть некие независимые признаки U и V принимают k и l равновероятностных значений[31]. Рассмотрим, каким свойствам должна удовлетворять функция f, характеризующая неопределенность распределений признаков.

$$F = f\ (k),$$

т.е. рассматриваемая функция зависит от числа градаций того признака, неопределенность распределения которого она измеряет и f (1) = 0. Для k > l должно быть справедливо неравенство

[31] Яглом А.М., Яглом И.М. Вероятность и информация. М.: Гос. изд-во физмат. литературы, 1960. С. 45.

$$f(k) > f(l).$$

Число сочетаний значений признаков равно произведению kl. Степень неопределенности двумерного распределения f (kl) должна быть равна сумме неопределенностей соответствующих одномерных распределений, т.е. f (kl) = f (k) + f (l). Логарифмическая функция – единственная функция аргумента k, удовлетворяющая условиям: f (k l) = f (k) + f (l), f (1) =0, f (k) >f (l) при k > l.

Энтропия может быть нулевой и ненулевой.

1. H (Y) = 0. Равенство достигается тогда, когда Y принимает только одно значение. Это ситуация максимальной определенности: случайным образом выбрав объект, мы точно знаем, что для него рассматриваемый признак принимает упомянутое значение (рис. 9). Единственная отличная от нуля вероятность здесь равна 1. Для такого распределения энтропия действительно равна нулю.

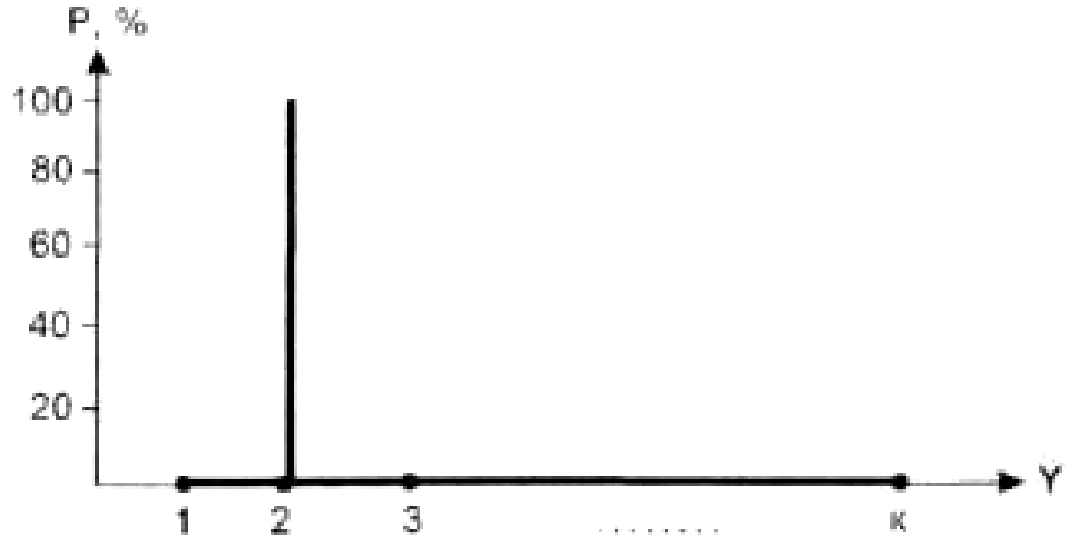


Рис. 9. Распределение с нулевой энтропией

2. При фиксированном k значение энтропии максимально, когда все возможные значения Y равновероятны. Это ситуация максимальной неопределенности. Предположим, что k = 5 (рис. 10).

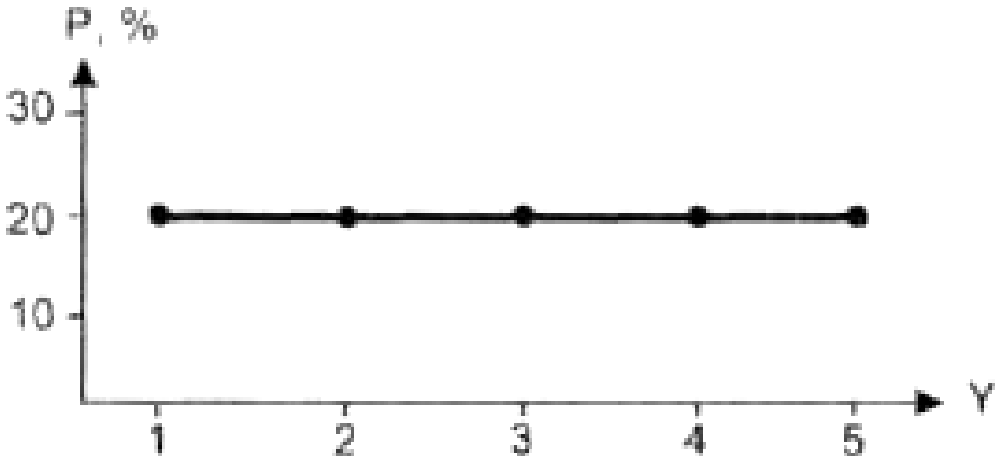


Рис. 10. Распределение с максимальной энтропией при заданном числе градаций признака

Здесь $P_j = 0{,}2$ и значение энтропии при этом равно log 5, а в общем случае в ситуации полной неопределенности энтропия равна log k. Чем больше градаций имеет признак, тем большей энтропии может достичь отвечающее ему распределение.

На рис. 9 и 10 мы видим 2 варианта: а) минимальная (нулевая) энтропия, наилучший прогноз, полная определенность; б) максимальная энтропия (равная log k и поэтому зависящая от числа градаций рассматриваемого признака), наихудший прогноз, полная неопределенность.

На рис. 9 разброс рассматриваемого признака равен нулю, а на рис.10 – максимально большой. Энтропия будет тем больше, чем реальное распределение ближе к ситуации, отраженной на рис. 10, и тем меньше, чем оно ближе к ситуации, отраженной на рис. 9. Поэтому энтропия может использоваться при оценке степени разброса значений номинального признака. Однако максимальное значение энтропии для распределения какого-либо признака зависит от числа его градаций. Энтропия не может выступать в качестве меры разброса – значение энтропии необходимо нормировать, поделить на величину максимальной энтропии и получить **энтропийный коэффициент**

$$\varepsilon = \frac{H}{H_{\max}} = \frac{H}{\log k}.$$

Тема 4
Типы шкал и методы анализа информации

В зависимости от того, насколько широк круг математических операций, допустимых для обработки и получения выводов, в социологии чаще всего используют шкалы следующих типов (если расположить их в порядке возрастания соответствующего *уровня измерений*)[32]: номинальные, ранговые, интервальные, пропорциональные. Все они разработаны и введены в научный оборот С. Стивенсом. Чем выше уровень шкалы, тем больше математических действий можно совершать с соответствующими числовыми значениями. Проблемы, которые возникают при построении одномерных частотных таблиц, связаны с типом шкалы.

4.1. Номинальная шкала

С помощью *номинальной шкалы* (шкалы наименований) мы измеряем такие переменные, которые не могут количественно отличаться друг от друга: каждое значение представляет собой отдельную категорию и является своего рода ярлыком или именем. Значения невозможно сравнивать между собою по принципу «больше-меньше», «выше-ниже» и т.п. Такие переменные невозможно складывать, вычитать, умножать и делить. Поэтому данные, полученные по номинальной шкале, резюмируются с помощью простого *частотного распределения (*табл. 7 и 8)[33].

Для данных номинального уровня измерение центральной тенденции производится с помощью определения *моды* (в табл. 7 модальную категорию представляют женщины, в табл. 8 – неработающие пенсионеры). Выявляя центральную тенденцию, следует обращать внимание на максимальные и минимальные зна-

[32] Добреньков В.И., Кравченко А.И.. Методы социологического исследования. С. 194.

[33] Там же.

чения изучаемой переменной – это сразу дает представление о масштабах изменения рассматриваемой переменной.

Таблица 7

Распределение респондентов по полу

Пол	Частота	Процент
Мужчины	399	44,3
Женщины	496	55,0
Всего	895	100,0

Таблица 8

Распределение респондентов по социально-профессиональному статусу

Социально-профессиональный статус	Частота	Процент
Руководители предприятий	16	1,8
Предприниматели	52	5,8
ИТР	83	9,3
Непроизводственная интеллигенция	89	9,9
Служащие без специального образования	48	5,4
Квалифицированные рабочие	93	10,4
Рабочие средней и низкой квалификации	102	11,4
Неработающие пенсионеры	226	25,3
Прочие	186	20,8
Всего	895	100,0

Помимо центральной тенденции измеряют и *дисперсию*. Для данных номинального уровня наибольшая дисперсия проявляется в тех случаях, когда наблюдения распределены поровну между категориями (например, одинаково число мужчин и женщин). Полное отсутствие дисперсии проявляется в тех случаях, когда все наблюдаемые значения переменной совершенно однородны – это представляет существенное препятствие для дальнейшего анализа. Например, при изучении взаимосвязи между полом и за-

нятостью в выборке опроса оказались одни мужчины. Поскольку налицо отсутствие дисперсии (т.е. нет вариаций по полу), сравнение провести нельзя. Самый простой одномерный анализ, проведенный в процессе сбора данных, поможет скорректировать выборку.

Удобным средством такого анализа служит графическое отображение рядов распределений. На рис. 11 в виде столбчатой диаграммы изображено распределение[34], представленное в табл. 8. Столбчатая (столбиковая) диаграмма представляет собой ряд столбцов; каждый из них – это процент или частота данного значения переменной. На рис. 12 приведена круговая (*pie-diagram* – «пирожковая диаграмма») диаграмма реестра голосов, поданных на выдвижении кандидатов в президенты[35].

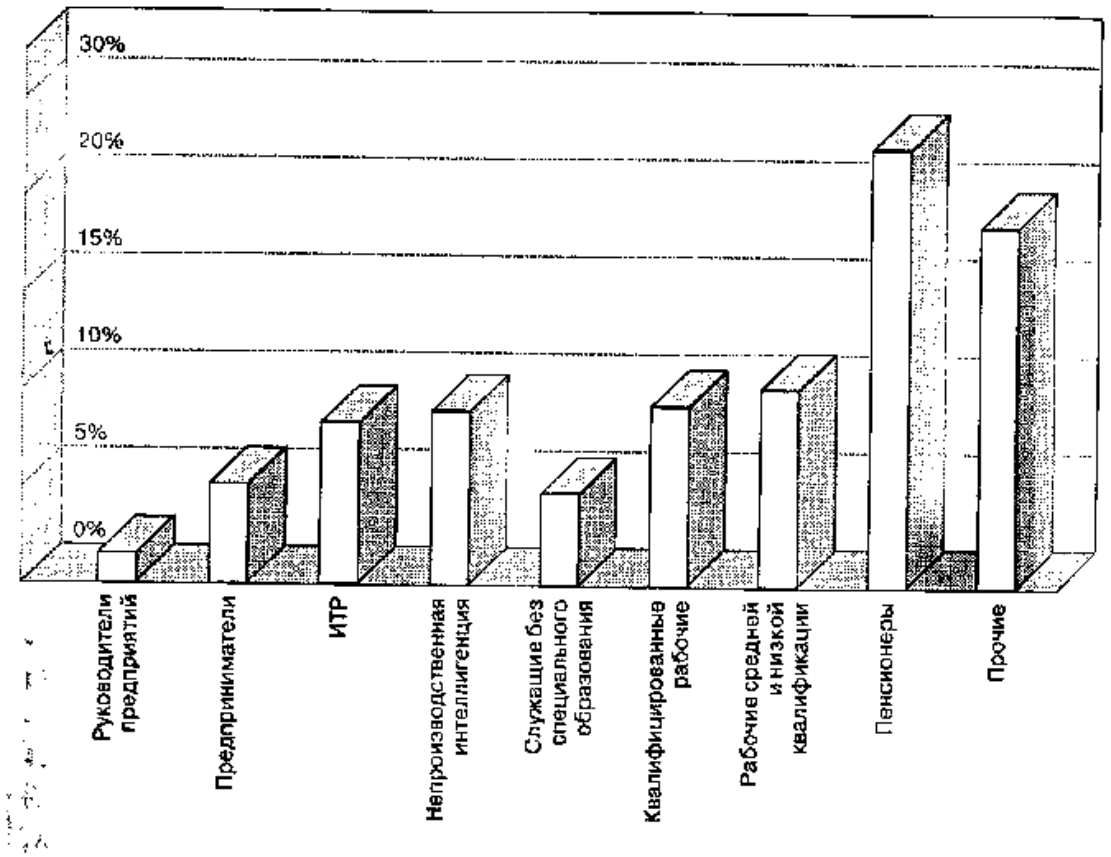


Рис. 11. Столбчатая диаграмма

[34] Добреньков В.И., Кравченко А.И. Указ. соч.
[35] Там же.

Рис. 12. Распределение голосов, поданных за кандидатов

Кумулята для номинальной шкалы не строится, полигон построить можно. Но отрезки, связывающие отдельные точки, мы не можем интерпретировать.

4.2. Ранговая шкала

Одномерная статистика, что используется для данных номинального уровня, может быть применена и для данных рангового уровня. Данные *рангового уровня* измерений включают в себя категории наблюдения, которые размещены по порядку (от большего значения признака к меньшему или наоборот, такие шкалы называют также порядковыми или ординальными.). Здесь мы можем выбрать для анализа:

– центральную тенденцию частотного распределения (моду или *медиану* – категорию, к которой принадлежит серединное наблюдение);

– разброс (дисперсию или среднеквадратическое отклонение).

Среднее отклонение (MD) представляет собой меру разброса, основанную на отклонении каждого из значений от среднего:

$$MD = \frac{\sum |x_i - \bar{x}|}{N}.$$

Если мы берем каждую отметку и вычитаем из нее среднее, мы вычисляем ту величину, на которую каждая из отметок отличается от среднего. Сумма этих отклонений всегда равна нулю. Мы не интересуемся знаком и находим *абсолютные значения* отклонения. Затем мы берем их сумму и делим на число отметок, чтобы найти среднее отклонение отметок от среднего. Чем боль-

ше среднее отклонение, тем сильнее разброс отметок вокруг среднего.

В табл. 9 значения переменных – частоты использования того или иного источника – соотнесены с ранговой шкалой, значения которой меняются от категории «часто» (ранг 4) до «не дали ответа» (ранг 0)[36]. Число наблюдений равно 426, половина наблюдений составит 213. Это означает, что медиана для такого источника информации, как «Встречи с мэром и работниками администрации», приходится на категорию с рангом I («никогда»); для четырех последующих переменных – на категорию с рангом 2 («иногда»); для последней переменной – «Телевидение» – медиана приходится на категорию 4 («часто»).

Таблица 9

Источники информации о работе городской администрации

Источники информации	Частота/ранг				
	часто	регулярно	иногда	никогда	не дали ответа
	4	3	2	1	0
Встречи с мэром и работниками администрации	2	5	39	282	98
Газеты	46	76	171	71	62
Общение с коллегами по работе	30	63	124	104	105
Общение с родными, соседями, друзьями	45	82	167	52	80
Радио	66	88	142	64	66
Телевидение	133	129	121	22	21

Кумуляту для порядковых шкал строить можно. Но интерпретация полигонов и гистограмм (и для кумуляты, и для выборочной оценки функции плотности распределения) может быть различной.

[36] Добреньков В.И., Кравченко А.И. Указ. соч. С. 202 – 203.

4.3. Интервальная шкала

Непрерывные интервальные шкалы не самые важные для социолога – даже возраст социологом часто рассматривается как номинальная или порядковая переменная: выделяются классы работающих и пенсионеров, молодежи и более старших людей, репродуктивный возраст и нерепродуктивный и т.д. Но они также часто используются.

Измерения интервального и пропорционального уровня редко анализируются с помощью прямого указания частот или процентных отношений. Значения переменных, измеряемых с помощью *интервальных шкал,* представляют собой численные величины, а не категории. При измерении доходов трудно рассчитывать, что суммы доходов различных респондентов или их семей будут совпадать до рублей и копеек. По этой причине значения таких переменных и размещают в *интервалах.*

Критериями центральной тенденции для пропорционального и интервального уровней измерений выступают мода, медиана и среднее арифметическое. *Среднее арифметическое* представляет собой сумму значений переменной, разделенную на число значений:

$$\overline{x} = \frac{\sum x_i}{N} = \frac{x_1 + x_2 + \ldots + x_i}{N},$$

где x_i – числовое значение *i*-й позиции, а *N* – объем выборки.

Рассмотрим вычисление средней арифметической величины на примере расчета средней посещаемости занятий в студенческой группе по данным проверок деканата (табл. 10)[37]. Сложив числа в правой колонке и разделив их на 10 (число проверок), мы получим, что средняя посещаемость в группе $x = 18{,}6$. Полученные средние величины следует *нормировать,* разделив их на численность студентов каждой группы.

Среднее может оказаться обманчивым показателем центральной тенденции, если среди значений переменной появится какая-то экстремальная величина. Медианный подход даст более корректные показатели. Если среднее и медиана различаются, то

[37] Добреньков В.И, Кравченко А.И. Указ. соч. С. 202 – 203.

предполагаем, что на значение среднего влияют одно или несколько экстремальных значений измеряемой переменной.

Таблица 10

Посещаемость занятий студентами группы

Номер занятия	Число присутствующих	Номер занятия	Число присутствующих
1	17	6	20
2	21	7	16
3	18	8	17
4	14	9	21
5	20	10	22

Для переменных, значения которых измеряются не однозначно определенными числами, а изменяются вдоль непрерывного ряда значений, рассчитывается не среднее арифметическое, а средневзвешенное. Предположим, что нам требуется вычислить средний возраст опрошенных респондентов (табл. 11)[38].

Таблица 11

Распределение респондентов по возрасту

Возраст, годы	Частота	%
18-24	46	10,1
25-29	55	12,0
30-39	97	21,2
40-49	115	25,2
50-59	74	16,2
60-70	70	15,3
	457	100,0

[38] Добреньков В.И, Кравченко А.И. Указ. соч.

Вначале мы должны определить середину каждого интервала путем вычисления простого среднего, т. е. сумма крайних значений делится пополам. Затем необходимо умножить это значение на число респондентов соответствующего возраста, сложить полученные произведения и разделить на общий объем выборки (см. табл. 11а)[39].

Таблица 11а

Результат 2-го этапа вычисления средневозрастной величины

Возраст, годы	Частота	Середина интервала	Произведение
18-24	46	21	966
25-29	55	27	1 485
30-39	97	34,5	3 346,5
40-49	115	44,5	5 117,5
50-59	74	54,5	4 033
60-70	70	65	4 550
Всего	457	2	19 498

Разделив полученную сумму на 457, мы получим средний возраст – 42,6 года. Формула для средневзвешенного значения выглядит аналогично с учетом того, что *x* здесь относится к середине интервала:

$$\overline{x} = \frac{\sum x_i n_i}{N} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_i n_i}{N},$$

где *x* – числовое значение некой позиции; n – число респондентов, наблюдаемых на данной позиции переменных; N – общее число наблюдений.

Показатели разброса данных интервального или пропорционального уровня включают среднее отклонение, дисперсию и среднеквадратическое отклонение.

Дисперсия – сумма квадратов отклонений от среднего, разделенная на число отметок:

[39] Добреньков В.И, Кравченко А.И. Указ. соч.

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

Среднеквадратическое отклонение представляет собой корень квадратный из дисперсии:

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

Чем больше разброс данных вокруг среднего, тем выше значения дисперсии и среднеквадратического отклонения. Это означает, что если все данные одинаковы, то S^2 равна нулю.

Таким образом, алгоритм для вычисления дисперсии и среднеквадратического отклонения таков[40]:

1. Вычислить среднее.
2. Вычислить разности между средним и каждым из значений.
3. Возвести в квадрат разности, вычисленные на этапе 2.
4. Умножить квадраты разностей на частоты наблюдений каждого из значений.
5. Просуммировать квадраты разностей, вычисленные на этапе 4.
6. Разделить сумму квадратов, полученную на этапе 5, на *N;* это равняется дисперсии.
7. Извлечь квадратный корень из числа, вычисленного на этапе 6; это равняется среднеквадратическому отклонению.

В зависимости от того, насколько велика (мала) дисперсия, или среднеквадратическое отклонение, мы можем судить, насколько единодушны были в своих оценках респонденты (при меньшем значении дисперсии), или насколько сильно они расходятся в своих мнениях (при большем значении дисперсии).

Интервальность шкалы обычно сопрягается с ее *непрерывностью*, т.е. в качестве значения интервального признака может выступить любое действительное число, любая точка числовой оси. А непрерывную кривую в выборочном исследовании нельзя получить никогда. Здесь мы не можем иметь линию, похожую на «колокол» нормального распределения. Даже если в генеральной

[40] Добреньков В.И, Кравченко А.И. Указ. соч.

совокупности распределение нормально, а выборка – репрезентативна, мы вместо «колокола» получим лишь некоторое его подобие, составленное, например, из отрезков, соединяющих отдельные точки – полигон распределения (рис. 13). Заменяющая непрерывное распределение ломаная линия может состоять также из «ступенек», в таком случае она называется гистограммой распределения (рис. 14).

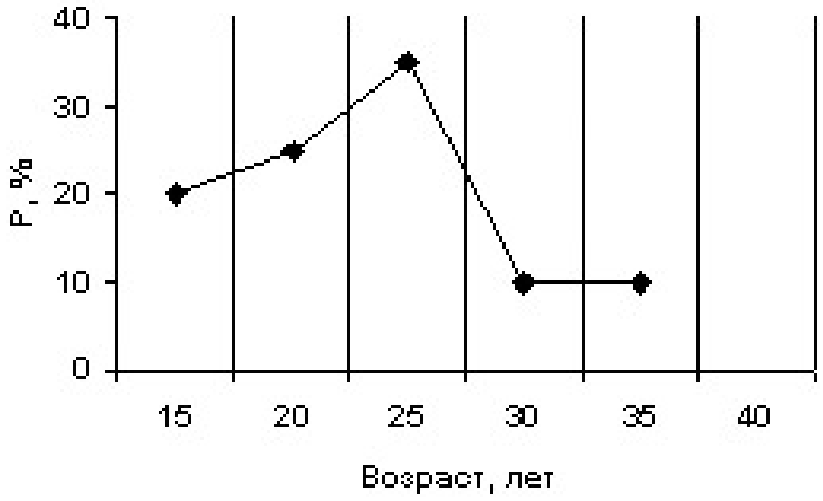


Рис 13. Полигон плотности распределения непрерывного признака

От середин отрезков, отмеченных на горизонтальной оси, откладываются проценты, соответственно 20, 25, 35, 10, 10.

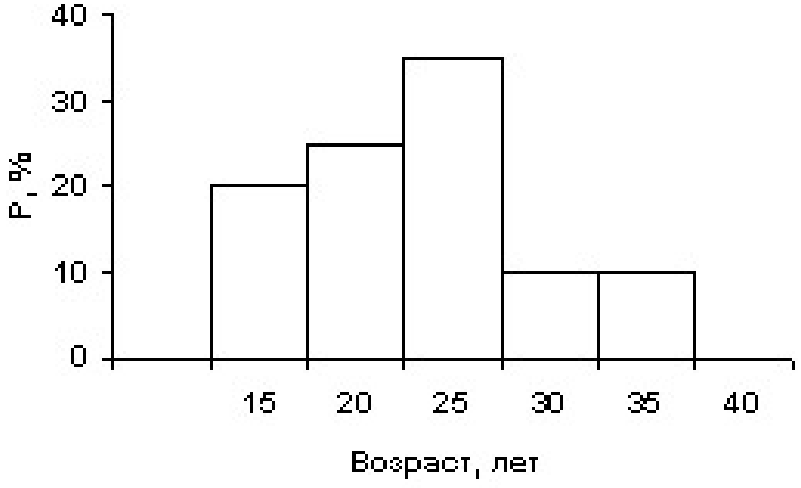


Рис. 14. Гистограмма плотности распределения непрерывного признака

При больших объемах выборки и достаточно мелком разбиении и гистограмма, и полигон хорошо отражают функцию плотности распределения (причем полигон делает это несколько лучше).

Для примера рассмотрим признак «возраст респондента». Рассмотрим два полигона распределения респондентов по возрасту. Первый полигон, при построении которого использовались все наблюденные значения возраста, изображенные на рис. 15, мы будем воспринимать как некий бессмысленный набор чисел. А если мы сгруппируем соответствующие наблюдения в интервалы 15 – 20 и 25 – 30 лет и приведем полигон к другому виду – виду, изображенному на рис. 16, то станет ясно, что изучаемая совокупность респондентов характеризуется тем, что половину ее составляют те, кто моложе 20 лет, а людей от 25 до 30 лет в ней вдвое меньше и т.д. Из таких фактов можно сделать содержательные выводы.

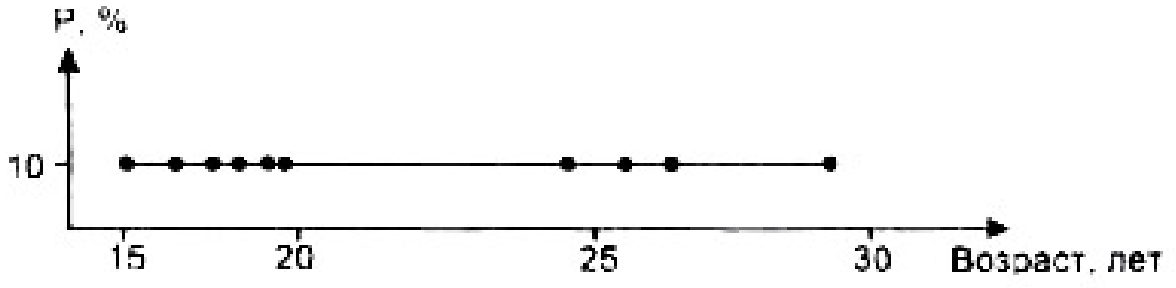


Рис. 15. Непродуктивный полигон распределения по возрасту

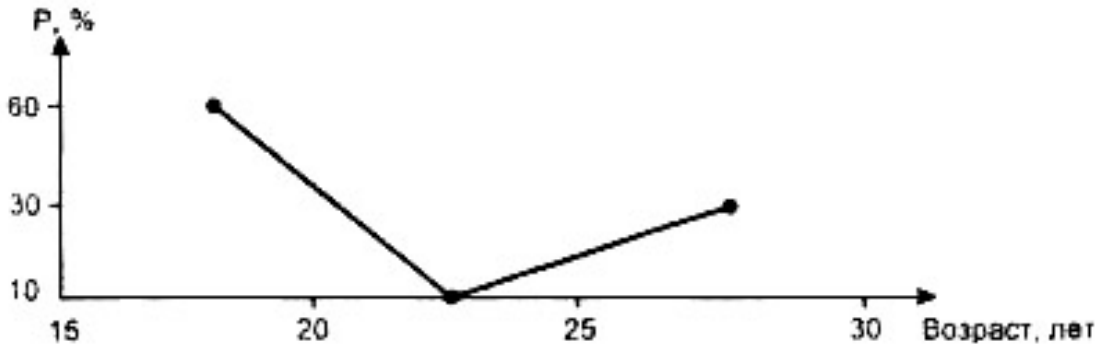


Рис. 16. Продуктивный полигон распределения по возрасту

Итак, мы получим закономерность, поскольку она позволяет нам сформировать какое-то новое представление об изучаемой совокупности респондентов – представление, связанное с описа-

нием совокупности «в среднем». Правда, здесь требуется подчеркнуть, что возможна двоякая интерпретация нашего шага.

1. Мы прибегли к определенному «сжатию» информации только потому, что не имели возможности прямо противоположного способа действий: скажем, измерения возраста с точностью до одного месяца и использования репрезентативной выборки в сотни тысяч единиц. Имея возможность сделать это, мы получили бы полигон, неотличимый на глаз от непрерывной кривой. Указанный подход, называемый обычно методом группировки, является более экономным способом записи информации, содержащейся в выборке (практически бесполезно знать 10 000 наблюдений, заданных на отрезке (0, 10), достаточно указать, какая доля наблюдений содержится в интервале (0, 1), (0, 2) и т.д.).

2. Даже если при дальнейшем дроблении величины интервалов распределение респондентов по возрасту будет стремиться к определенному виду, этот вид может вообще не интересовать социолога. Многие числовые характеристики людей (например, возраст), чаще всего интересуют социолога не сами по себе, а как признаки – индикаторы, чего-то латентного (возраст служит для оценки социальной зрелости опрашиваемого).

Кроме того, мы должны «сжать» исходные данные путем разбиения диапазона изменения значений этого признака на интервалы. За счет потери одной информации мы приобретаем другую.

Предположим, что мы изучаем связь между двумя признаками[41]: Y, принимающим два значения – 1 и 2, и X, принимающим четыре значения – 1, 2, 3, 4 (табл. 12).

Между X и Y имеется статистическая связь. Если бы связи не было, то внутри каждого значения признака X респонденты должны были бы поровну распределяться между двумя категориями признака Y (первая строка должна была бы состоять из частот 25 и 25, вторая – 24 и 24, третья – 21 и 21, четвертая – 20 и 20).

Предположим теперь, что мы сгруппировали значения признака X, объединив градации 1 и 2, градации 3 и 4, т.е. разбили значения признака X на интервалы (табл. 13). Связь между искомыми признаками не фиксируется.

[41] Миркин Б.Г. Анализ качественных признаков и структур. М.: Статистика, 1980. С. 18.

Таблица 12

Сопряженность при наличии связи между признаками X и Y

Значения X	Значения Y		Итого
	1	2	
1	44	6	50
2	5	43	48
3	38	4	42
4	3	37	40
Итого	90	90	180

Таблица 13

Объединение градаций (1 и 2) и (3 и 4) признака X

Значения X	Значения Y		Итого
	1	2	
1+2	49	49	98
3+4	41	41	82
Итого	90	90	180

Сгруппируем значения признака X по-другому, т.е. разобьем совокупность этих значений на иные интервалы: объединим градации 1 и 3, а также градации 2 и 4 (табл. 14). Здесь мы фиксируем наличие связи.

Таблица 14

Объединения градаций (1 и 3) и (2 и 4) признака X

Значения X	Значения Y		Итого
	1	2	
1+3	82	10	92
2+4	8	80	88
Итого	90	90	180

При определении способа разбиения диапазона изменения признака на интервалы мы должны ориентироваться:

– на задачу исследования. Так, при изучении типов личности, вполне возможно, что нас удовлетворит разбиение всех возрастов от 15 до 100 лет на равные интервалы: (15 – 20), (20 – 25),

(25 – 30) и т.д. Если же одной из решаемых нами задач будет изучение выбора молодежью жизненного пути, то отдельно рассмотрим интервалы (15 – 17) – в 17 лет человек заканчивает школу; (17 – 18) – в 18 лет юношей забирают в армию; (18 – 22) – в 22 года большинство поступивших в институт получают дипломы и т.д.[42];

– на возможность сравнивать свои результаты с результатами других социологов (способы разбиения диапазонов изменения тех признаков, по которым совокупности сравниваются, должны быть одинаковыми).

При этом возникают следующие **проблемы:**

1. *Значение рассматриваемого признака лежит на стыке двух интервалов.* Все стыки считают принадлежащими правому интервалу: будем рассматривать полуинтервалы: [15, 20), [20, 25) и т.д. Последним полуинтервалом может быть, например, [60, 65). Правый конец самого правого интервала можно увидеть так: вместо полуинтервала [60, 65) использовать отрезок [60, 65]; ввести дополнительный полуинтервал [65, 70).

2. *Проблемы построения полигонов, гистограмм.* Вертикаль, на которой будет откладываться величина процента при построении полигона, может начинаться в любой точке интервала (хотя на практике чаще используют середину).

3. *Проблема выбора графического изображения с большей наглядностью.* Обычно считают, что полигон отвечает кусочно-линейной плотности распределения. При использовании же гистограммы полагают, что объекты равномерно распределены внутри каждого интервала. В соответствии с теорией вероятностей площадь фигуры, лежащей под кривой функции плотности над каким-либо интервалом, равна вероятности попадания объекта в этот интервал. В случае гистограммы вероятность попадания равна площади соответствующего отрезку прямоугольника гистограммы.

4. *Проблема возникновения гистограммы с неравными интервалами.* Например, мы интересуемся категориями людей, с

[42] См. подробнее: Сиськов В.И. Об определении величины интервалов при группировках // Вестник статистики. 1971. № 12; Пасхавер Б. Проблема интервалов в группировках // Вестник статистики. 1972. № 6.

одной стороны, думающих о вступлении в фазу трудовой деятельности и вступающих в нее (15 – 20 лет), с другой стороны – собирающихся покинуть эту фазу (50 – 55 лет). Предположим, что частотная таблица, на базе которой мы хотим построить гистограмму, имеет вид, отраженный в табл. 15.

Таблица 15

Частотное распределение респондентов по возрасту

Интервал изменения возраста	[15 – 20)	[20 – 50)	[50 – 55)	[55 – 80)
Количество респондентов, попавших в интервал	80	90	20	10

На основе данных табл. 15 вычерчиваем график (рис. 17):

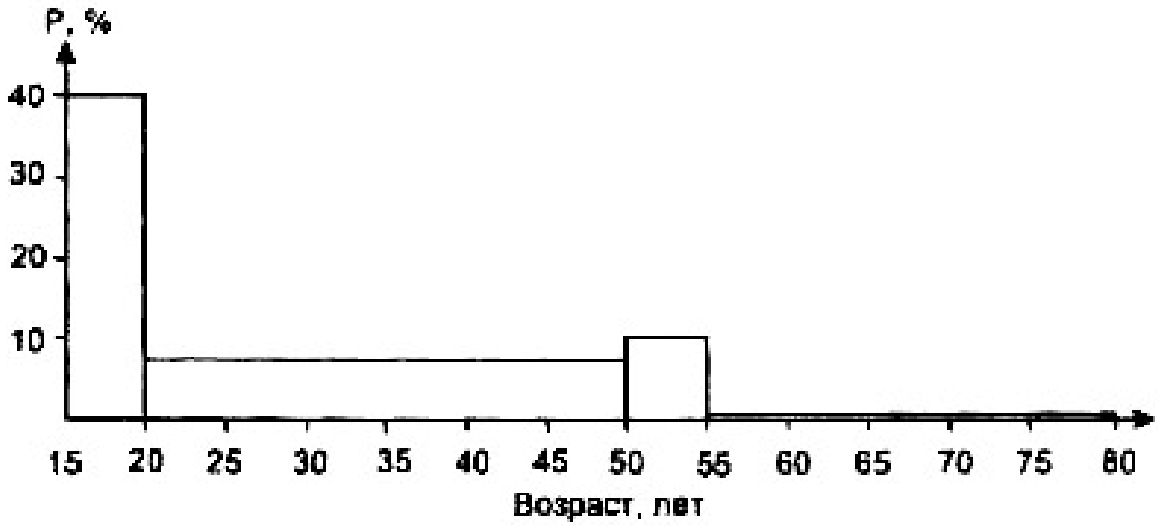


Рис. 17. Гистограмма, построенная на основе табл. 15.

Выбираем какой-то интервал диапазона изменения возраста за единицу и считаем, что на нем высота столбца гистограммы равна проценту людей, попавших в этот интервал. Искомые совокупности отражены в интервалах [15 – 20) и [50 – 55). Другими словами, мы выбрали за единицу интервал длиной в 5 лет. Для интервалов, имеющих другую длину, высоту столбца гистограммы будем полагать равной результату деления величины процента попавших в него людей на длину интервала. Так, интервал [50 – 55) имеет длину в 6 единиц. В него попали

45% респондентов. Поделим 45 на 6. Получится 7,5%. Именно такой высоты столбец и будет отвечать рассматриваемому интервалу. Так же поступим с интервалом [55 – 80). В него попало 5% респондентов, а длина его равна 5 единицам. Значит, высота соответствующего столбца равна 50: 5 = 1 %.

Площадь каждого столбца будет равной проценту респондентов, возраст которых попал в интервал, лежащий в его основании. Только тогда гистограммы, представляющие функцию плотности нормального распределения, будут напоминать форму «колокола» и при увеличении дробности интервалов приближаться к «гладкой» кривой.

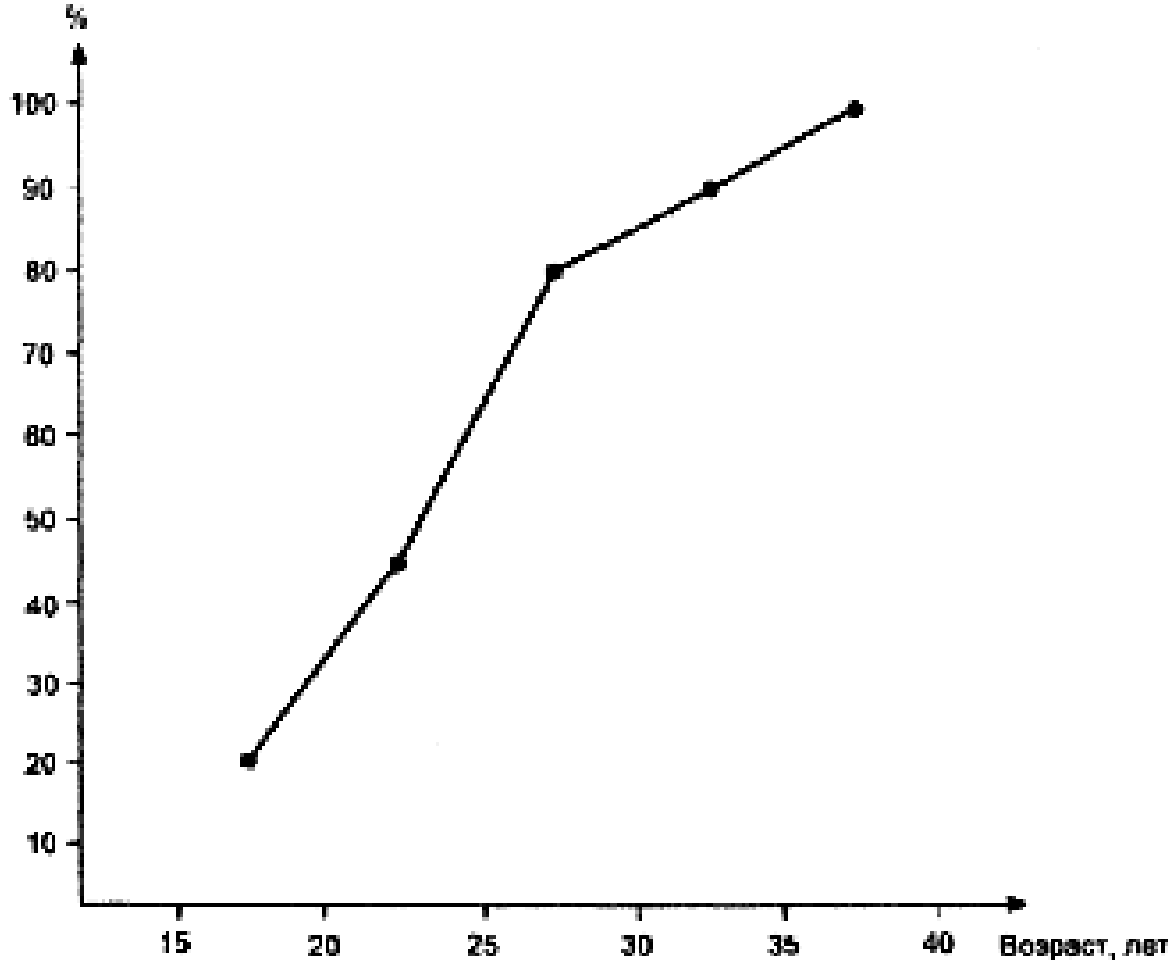


Рис. 18. Кумулята распределения, отвечающего выборочной функции плотности

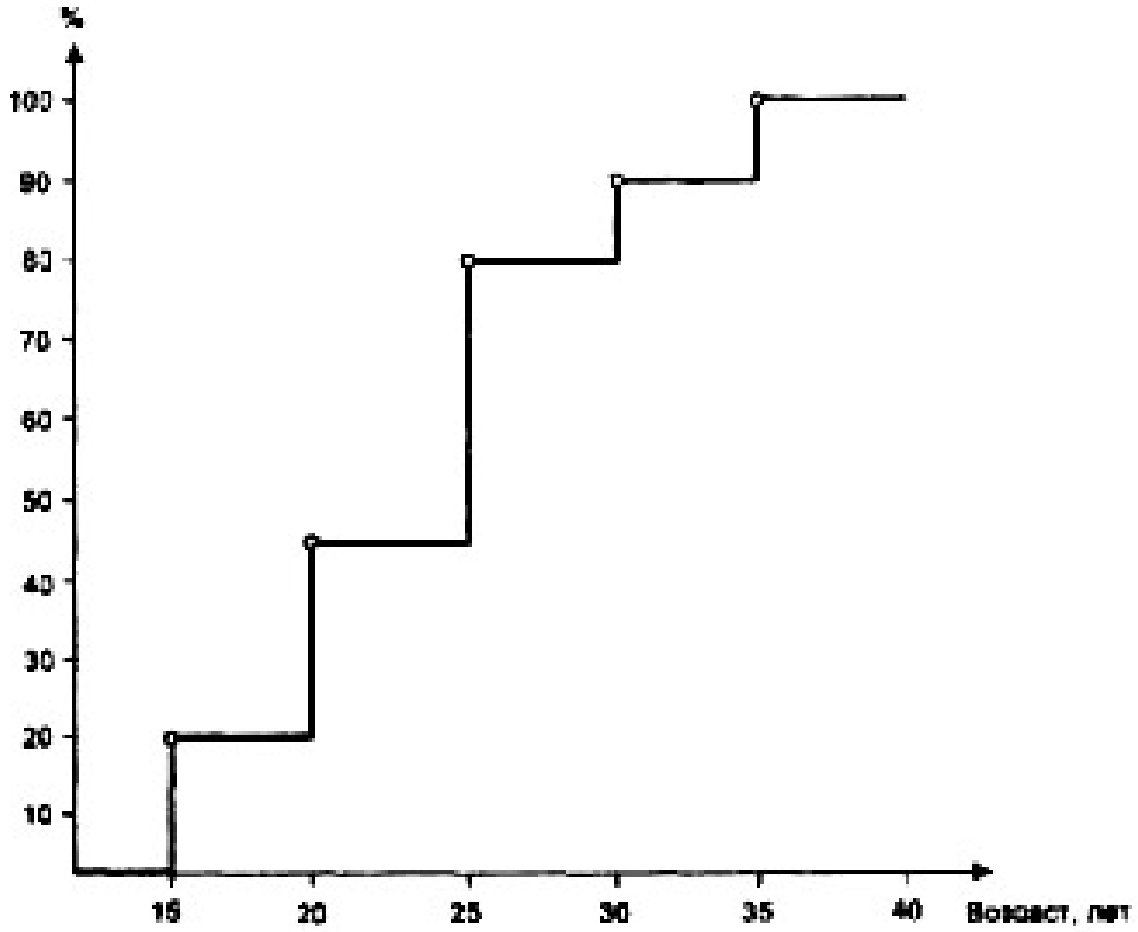


Рис. 19. Кумулята распределения, заданная в виде гистограммы

Выборочным представлением собственно функции распределения случайной величины, стоящей за рассматриваемым признаком, служит т.н. **кумулята** распределения, или график накопленных частот. Она обычно представляется в виде полигона, каждая вершина которого отвечает относительной частоте того, что признак принимает значение, не превышающее того, над которым эта вершина находится. Кумулята получается из описанного выше полигона распределения путем последовательного суммирования определяющих его частот. Так, полигону, изображенному на рис. 17, будет отвечать следующая кумулята (рис.18): полуинтервалу (25, 30] соответствует частота 80%, складывающаяся из частот, соответствующих полуинтервалам (15, 20], (20, 25] и (25, 30]. Выборочное представление функции распределения может быть задано и в виде гистограммы (рис. 19).

Тема 5
Анализ двухмерных распределений

5.1. Общая характеристика двухмерных распределений

В гипотезе исследования, как правило, высказывается предположение о наличии связи между двумя и более переменными. Чтобы анализировать этот аспект, необходимо найти ответ на пять основных вопросов[43]:

1. Существует ли обозначенная в гипотезе связь между независимой и зависимой переменными?
2. Каково направление этой связи?
3. Насколько сильна связь?
4. Является ли связь статистически значимой?
5. Является ли связь каузальной?

Мы можем утверждать, что связь существует, если наблюдаемые значения независимой переменной ассоциируются с наблюдаемыми значениями зависимой переменной. Выдвинем гипотезу: «Чем старше избиратели, тем больше вероятность того, что они примут участие в выборах». При анкетировании задаем прямой вопрос с предлагаемыми вариантами ответов:

Принимали ли Вы участие в последних выборах главы городского самоуправления?

1 – да;

2 – нет;

3 – не помню.

При обработке данных опроса нам для проверки гипотезы необходимо сопоставить значения независимой переменной (возраст) с соответствующими им значениями зависимой переменной (участие или неучастие в выборах). С целью такого сопоставле-

[43]. Добреньков В.И., Кравченко А.И. Методы социологического исследования. С. 210.

ния мы после соответствующей обработки данных составляем табл. 16.

Таблица 16

Участие в выборах избирателей различных возрастов

Возраст, лет	Участие в голосовании, %				Всего
	Нет ответа	Да	Нет	Не помнят	
18-24	0	16	27	3	46
Процент по строке	0	34,8	58,7	6,5	9,2
Процент по столбцу	0	5,3	17,4	7,9	
25-29	0	30	18	7	55
Процент по строке	0	54,5	32,7	12,7	11,0
Процент по столбцу	0	10,0	11,6	18,4	
30-39	3	58	27	9	97
Процент по строке	3,1	59,8	27.8	9,3	19,4
Процент по столбцу	50,0	19,3	17,4	23,7	
40 – 49	1	75	32	7	115
Процент по строке	0.9	65,2	27,8	6,1	23,0
Процент по столбцу	16,7	24.9	20,6	18,4	
50 – 59	0	48	20	6	74
Процент по строке	0	64,9	27,0	8,1	14,8
Процент по столбцу	0	15,9	12,9	15,8	
60-70	0	49	18	3	70
Процент по строке	0	70,0	25,7	4,3	14,0
Процент по столбцу	0	16,3	11,6	7,9	
Старше 70	2	25	13	3	43
Процент по строке	4,7	58.1	30,2	7,0	8,6
Процент по столбцу	33,3	8,3	8,4	7,9	
Всего	6	301	155	38	500
Процент	1,2	60,2	31,0	7,6	100,0

Такая таблица называется *«кросстаб»*[44], а процесс ее создания – *кросстабуляция.* Двигаясь по строкам, мы начинаем с первого значения независимой переменной (возраст) 18 – 24 года. Мы видим, что здесь число принимавших участие в выборах заметно меньше числа тех, кто не участвовал. Перейдя к следую-

[44] Добреньков В.И., Кравченко А.И. Указ. соч.

щей строке, 25 – 29 лет, мы видим, что здесь соотношение между числом участвовавших и не участвовавших противоположное: первых уже в два с лишним раза больше. Это соотношение еще более возрастает при переходе к следующим возрастным категориям, хотя и несколько снижается для самой старшей группы избирателей (старше 70 лет). Это позволяет нам сделать выводы:

а) о *наличии связи* между независимой (возраст) и зависимой (участие в выборах) переменными;

б) о *направлении* этой связи, которая в данном случае является прямой или положительной, поскольку ее можно выразить следующим простым описанием: *чем больше* значения независимой переменной (возраст), *тем больше* значения зависимой переменной (процент участия в выборах). Исключение составляет лишь самая верхняя возрастная группа, где электоральная активность по вполне понятным причинам снижается.

Когда низкие значения одной переменной ассоциируются с низкими значениями другой переменной (и наоборот), имеет место положительная связь. Когда низкие значения одной переменной ассоциируются с высокими значениями другой, между двумя переменными существует отрицательная связь.

Иногда для большей наглядности анализа используют различные *индексы* – специально создаваемые показатели, с помощью которых связь между переменными проявляется более отчетливо (например, индекс электорального участия, равный частному от деления числа принимавших участие в каждой из возрастных групп на число тех, кто не голосовал; он также исчисляется не делением, а вычитанием).

Таким образом, мы фиксируем:

а) наличие связи (нет изменения – нет связи);

б) силу связи (насколько существенно различаются наблюдаемые значения зависимой переменной при изменении значений независимой переменной).

Наиболее сильная из возможных связей между двумя переменными – это такая связь, при которой значение зависимой переменной для каждого случая в одной категории независимой переменной отличается от каждого из случаев в другой категории (совершенная связь). Совершенная связь между независимой и

зависимой переменными дает исследователю возможность точно предсказать значение любого из случаев зависимой переменной, если известно значение независимой. Пример совершенной связи для гипотетического случая различий в голосовании приведен в табл. 17. Между переменными может существовать как совершенная положительная, так и совершенная отрицательная связь.

В реальных распределениях социологических данных крайне редко встречаются как совершенная связь, так и абсолютно полное ее отсутствие. Фактически отсутствие связи выражается в ее слабости. Слабой можно считать такую связь, при которой различия наблюдаемых значений зависимой переменной для различных категорий независимой переменной незначительны. Фактически наиболее слабая связь – это такая, в которой распределение было бы идентично для всех категорий независимой переменной (связь отсутствует).

Таблица 17

Различия в голосовании за кандидатов в зависимости от пола избирателей

Кандидат	Голосование, %	
	Мужчины	Женщины
Иванов	100	**0**
Петров	**0**	100
Всего	100	100

5.2. Показатели связи в двухмерных распределениях

Довольно часто используемым показателем силы связи выступают различные **коэффициенты корреляции**[45]. Корреляция указывает на степень статистической взаимосвязи признаков. Одним из индексов такого рода при использовании порядковой

[45] Яшин В.П. Корреляционный анализ в социологических и психологических исследованиях. Н.Новгород: Изд-во НКИ, 1999.

шкалы измерения выступает *коэффициент ранговой корреляции Спирмена*. Формула расчета его имеет следующий вид:

$$r_s = 1 - \frac{6\sum d_i^1}{l^3 - l}.$$

Коэффициент ранговой корреляции Спирмена будет равен +1 (абсолютная положительная связь), если ответы респондентов анализируемых групп будут в точности совпадать; он будет равен -1 (абсолютная отрицательная связь), если ответы всех респондентов обеих анализируемых групп будут прямо противоположны; если $r_s = 0$, то это означает полное отсутствие всякой связи. Коэффициент ранговой корреляции показывает, насколько одинаковыми или различными оказываются ответы на один и тот же вопрос со стороны двух сравниваемых между собою групп респондентов.

Тема 6
Анализ связей между номинальными признаками

6.1. Общая характеристика подходов к анализу номинальных данных

Роль номинальных данных в социологии огромна, что объясняется следующими причинами:

– простота их получения и естественность интерпретации;

– они более надежны, чем данные, полученные по шкалам более высокого типа;

– в методах, используемых для анализа номинальных данных, обычно бывают заложены модели, отвечающие естественной логике социолога.

Изучение связей между переменными, как правило, интересует исследователя не само по себе, а как отражение соответствующих причинно-следственных отношений. Однако социолог может наблюдать только статистические связи, а понятия «причина» и «следствие» не могут быть формализованы. Математика не может доказать, что такой-то признак служит причиной (следствием) того или иного явления. Для оценки связей между признаками используются частотные таблицы, или **таблицы сопряженности** (выборочные оценки вероятностных распределений многомерных случайных величин). На основе анализа подобных таблиц можно судить о сопряженности (совместной встречаемости) каких-то значений одних признаков с некоторыми значениями других признаков.

Предположим, что мы имеем два признака X и Y, первый из которых принимает значения 1, 2, ..., r, а второй – значения 1, 2, ..., c. Назовем двухмерной таблицей сопряженности 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):

$$\left\| n_{ij} \right\| = \begin{Vmatrix} n_{11} & n_{12} & \dots & n_{1c} \\ n_{21} & n_{21} & \dots & n_{2c} \\ \dots & \dots & \dots & \dots \\ n_{y1} & n_{y2} & \dots & n_{yc} \end{Vmatrix}.$$

Обычно ее представляют с явно обозначенными наименованиями признаков и их значений и выписанными маргинальными суммами:

Таблица 18

Общий вид таблицы сопряженности

Значения X	Значения 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). Нижняя строка образуется столбцовыми маргинальными суммами (маргиналами по столбцам). Объем выборки n равен сумме маргиналов по столбцам (либо по строкам). В качестве элементов таблицы могут фигурировать частоты, средние значения, мода, медиана, дисперсии, величины отклонений от средних по строке (столбцу), разница между эмпирической и теоретической частотой (пакет SPSS).

Действия исследователя могут быть направлены на:

– поиск сочетаний значений признаков, детерминирующих поведение респондента, т.е. поиск взаимодействий;

– поиск латентной переменной для каждого респондента.

Часто при этом используется т. н. *оцифровка*, т.е. приписывание каждой градации любого номинального признака определенного числа, отвечающего искомому значению соответствующей латентной переменной.

Математические методы работы с таблицами номинальных данных направлены на решение задач, типологизированных по классам. Рассмотрим классификацию задач, приведенную Ю.Н. Толстовой[46]. Ею выделяются следующие типы задач:

– *альтернатива – альтернатива,* т.е. такие, которые позволяют изучать связь между отдельными значениями любых рассматриваемых признаков (детерминационный анализ)[47];

– *группа альтернатив – группа альтернатив,* т.е. включающие анализ фрагментов таблиц сопряженности, методы выявления логических закономерностей, методы поиска детерминирующих сочетаний значений рассматриваемых признаков, в т.ч. алгоритмы, для обозначения которых используются аббревиатуры, включающие в себя сочетание AID (automatic interaction detector): CHAID, THAID[48];

– *признак – признак* (наиболее знакомые социологу коэффициенты парной связи);

– *признак – группа признаков*, включающие регрессионный анализ, методы построения индексов;

– *группа признаков – группа признаков,* предполагающие канонический анализ, или анализ соответствий, который дает возможность осуществлять оцифровку, изучать связи между признаками с т.н. «совместными» альтернативами, находить веса признаков при формировании индекса[49].

Тип задач, отвечающих рассмотрению всей совокупности признаков как системы, называется *анализом системы признаков* (логлинейный анализ, или причинный анализ).

---

[46] Толстова Ю.Н. Анализ социологических данных. С. 98.

[47] Чесноков С.В. Детерминационный анализ социально-экономических данных. М.: Наука, 1982. С. 276.

[48] Интерпретация и анализ данных в социологических исследованиях. М.: Наука, 1987. С. 136 – 151.

[49] Clausen S.-E. Applied correspondence analysis. An introduction. Sage university paper series on Quantitative applications in the social sciences, 07-121. Newbury park, CA: Sage, 1998.

6.2. Анализ связей типа «признак – признак»

Для измерения связи между двумя номинальными признаками предлагается более сотни коэффициентов. Мы рассмотрим лишь наиболее часто применяемые.

6.2.1. Коэффициенты связи, основанные на критерии хи-квадрат

Предположим, мы ищем зависимость профессии Y респондента от его пола X. Пусть анкета содержит соответствующие вопросы и в ней перечисляются пять вариантов профессий, закодированных цифрами от 1 до 5, для обозначения мужчин и женщин используются коды 1 и 2 соответственно, а исходная таблица сопряженности для 100 респондентов имеет вид:

Таблица 19

Сопряженность признаков «пол-профессия»

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

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) 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 – для случая пропорциональности строк: они будут пропорциональны и строке маргинальных сумм по столбцам.

Приведенная частотная таблица является результатом изучения выборочной совокупности респондентов. Но нас интересует не выборка, а генеральная совокупность, выборка же однозначно будет содержать т.н. выборочную ошибку. Учитывая это, мы 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, табл. 20 (по сравнению с табл. 19 в ней четыре частоты изменены на единицу) и табл. 21 (те же частоты изменены на две единицы). Таблица же 22 отличается от них.

Таблица 20

Сопряженность, 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 21

Сопряженность, частоты которой сравнительно 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 22

Сопряженность, 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

Сильное отклонение от пропорциональности заставляет нас сомневаться в отсутствии связи в генеральной совокупности; слабое отклонение говорит о том, что выборка не дает оснований для таких сомнений.

На основе функции *хи-квадрат* мы можем проверить гипотезу об **отсутствии связи.**

Предположим, что мы имеем две номинальных переменных, отвечающую им частотную таблицу и хотим определить, имеется ли связь между переменными, с помощью проверки статистической гипотезы о независимости признаков (суть нуль-гипотезы H_0 состоит в том, что связь между рассматриваемыми переменными отсутствует).

Допустим, мы хотим проверить статистическую гипотезу H_0[50]. Сделаем это с помощью числовой функции **f** от наблюдаемых величин, например, рассчитанной на основе частот выборочной таблицы сопряженности: $\mathbf{f = f\ (n_{ij})}$. Значение этой функции мы можем вычислить для нескольких выборок. 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: если H_0 справедлива, то для каждого полученного по конкретной выборке значения **f** можно сказать, какова та вероятность, с которой мы могли на него выбрать. Вычисляем

[50] Толстова Ю.Н. Анализ социологических данных. С. 102 – 110.

значение $\mathbf{f}_{выб}$ критерия **f** для нашей единственной выборки. Находим вероятность $\mathbf{P(f}_{выб}\mathbf{)}$ этого значения. Далее 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.

Если вероятность события $\mathbf{P\ (f}_{выб}\mathbf{)}$ очень мала, мы полагаем, что неправильно ее определили. Таким образом, наша гипотеза не подтверждается, т.к. мы изначально исходили из ее верности.

Если же вероятность $\mathbf{P\ (f}_{выб}\mathbf{)}$ достаточно велика для того, чтобы значение $\mathbf{f}_{выб}$ могло встретиться практически, то мы принимаем гипотезу: считаем, что она справедлива для генеральной совокупности.

Граница между малой и большой вероятностью должна быть равна такому значению вероятности, относительно которого мы могли бы считать, что событие с такой (или с меньшей) вероятностью практически не может случиться. Это значение называют уровнем значимости принятия (отвержения) нульгипотезы и обозначают буквой а. Обычно полагают, что а равно 0,05 либо 0,01.

Теперь рассмотрим гипотезу об отсутствии связи между двумя изучаемыми номинальными переменными. Функция, выступающая в качестве описанного выше статистического критерия носит название *хи-квадрат*. В разных случаях она обозначается большой или малой греческой «хи».

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

где $n_{ij}^{эмп}$ – наблюдаемая нами частота, стоящая на пересечении i-й строки и j-го столбца таблицы сопряженности (т.н. эмпирическая частота), а $n_{ij}^{теор}$ – частота, которая стояла бы в той же клетке, если бы наши переменные были статистически независимы (т.е. частота, отвечающая пропорциональности столбцов (строк) таблицы сопряженности; она называется теоретической, или ожидаемой частотой, поскольку именно ее появление и ожидается

при независимости переменных). Теоретическая частота находится по формуле:

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}.$$

Теоретическая частота отвечает той ситуации, когда являются независимыми два события:

а) первый признак принимает значение i;

б) второй признак принимает значение j.

Независимость двух событий означает, что вероятность их совместного осуществления равна произведению вероятностей осуществления каждого в отдельности. Эти вероятности оцениваются следующим образом:

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

Независимость наших событий означает справедливость соотношения:

$$P(X = i.Y = j) = P(X = i) \times P(Y = j).$$

Теперь рассмотрим работу критерия хи-квадрат. Представим себе, что мы организуем бесконечное количество выборок и для каждой из них вычисляем величину χ^2. Образуется последовательность таких величин: $\chi^2_{выб1}$, $\chi^2_{выб2}$, $\chi^2_{выб3}$ … Рассмотрим их распределение, т. е. вероятность встречаемости каждого значения. В математической статистике доказано следующее положение: если наши признаки в генеральной совокупности независимы, то вычисленные для выборок значения χ^2 приблизительно имеют хорошо изученное распределение с$^{2.}$ Приблизительность можно игнорировать, если в каждой клетке таблицы есть по крайней мере 5 наблюдений.

При отсутствии связи в генеральной совокупности среди выборочных χ^2 будут преобладать значения, близкие к нулю, поскольку отсутствие связи означает равенство эмпирических и теоретических частот. Большие значения χ^2 будут встречаться редко – именно они будут маловероятны. Поэтому можно ска-

зать, что **большое значение χ^2 приводит нас к утверждению о наличии связи, малое – об ее отсутствии.**

Вероятность попадания каждого значения величины в любой заданный интервал определяется с помощью специальных вероятностных таблиц. Такие таблицы имеются и для распределения c^2. В зависимости от вида таблицы типологизированы и сами эти распределения. Вид их определяется числом степеней свободы df (degree freedom) распределения:

$$Df = (r - 1)\,(c - 1).$$

Если в генеральной совокупности признаки независимы, то, вычислив df, мы можем найти по соответствующей таблице вероятность попадания произвольного значения χ^2 в любой заданный интервал. Вычисленное для нашей выборки значение обозначим $\chi^2_{выб}$.

Вычислим число степеней свободы df и зададимся некоторым уровнем значимости a. Найдем по таблице распределения c^2 такое значение $\chi^2_{табл}$, называемое критическим значением критерия $(\chi^2_{крит})$, для которого выполняется неравенство:

$$P(x \mathrm{I} \chi^2_{табл}) \neq a,$$

где x – обозначение случайной величины, имеющей распределение c^2 с рассматриваемым df.

Если $\chi^2_{выб} < \chi^2_{табл}$ (т. е. вероятность появления $\chi^2_{выб}$ достаточно велика), полагаем, что наши выборочные наблюдения не дают оснований сомневаться в том, что в генеральной совокупности признаки действительно независимы. Следовательно, мы принимаем нуль-гипотезу. Если $\chi^2_{выб}$ не равно $\chi^2_{табл}$ (т.е. вероятность появления $\chi^2_{выб}$ очень мала, т.е. меньше a), то мы отвергаем нуль-гипотезу – полагаем, что признаки зависимы.

В заключение следует отметить необходимость нормировки значений функции хи-квадрат. Сами значения рассматриваемого

критерия непригодны для оценки связи между признаками, поскольку они зависят от объема выборки и других случайных обстоятельств. Например, величина критерия 30 может говорить о большой вероятности наличия связи, если в клетках исходной частотной таблицы стоят величины порядка 10, 20, 30, и о малой вероятности того же, если рассматриваемые частоты равны 1 000, 2 000, 3 000 и т.д. **Социологу 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.** Нормировку осуществляют таким образом, чтобы нормированные коэффициенты изменялись либо от –1 до +1 (если выясняем положительную и отрицательную направленность), либо от 0 до 1 (во всех других случаях).

Имеются разные подходы к требующейся нормировке. Наиболее известными являются такие, которые превращают критерий хи-квадрат в известные коэффициенты – Пирсона (Р), Чупрова (Т), Крамера (К) соответственно:

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

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

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

Все коэффициенты изменяются от 0 до 1 и равны нулю в случае полной независимости признаков. Но с их помощью нельзя выделить зависимую и независимую переменные.

Обычно в качестве недостатка коэффициента Пирсона Р упоминается зависимость его максимальной величины от размера таблицы сопряженности (максимум Р достигается при с = r, но величина максимального значения изменяется с изменением числа категорий: при с = 3 значение Р не может быть больше 0,8; при с = 5

максимальное значение Р равно 0,89 и т.д.)[51]. Это приводит к возникновению трудностей при сравнении таблиц разного размера.

Для исправления этого недостатка коэффициента Пирсона Чупров ввел коэффициент Т. Но и Т достигает 1 лишь при c = r, и не достигает 1 при разном значении c и r. Может достигать 1 независимо от вида таблицы коэффициент Крамера К. Для квадратных таблиц коэффициенты Крамера и Чупрова совпадают, в остальных случаях К >Т[52].

6.2.2. Коэффициенты связи, основанные на моделях прогноза

Чтобы признаки считались связанными, значение одного из них должно позволять достаточно хорошо предсказать значение другого.

Предположим, что мы изучаем жителей некоторого города N от 20 лет и старше – нас интересует связь между признаком «возраст», рассматриваемым нами как номинальный, и дихотомическим признаком со значениями «студент – не студент». Предположим, что распределение изучаемой совокупности по возрасту приблизительно равномерно – такое, как изображено на рис. 20.

Мы не сможем хорошо прогнозировать возраст респондента. Выбрав наугад 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. Другое дело, если мы рассмотрим только студентов. Их распределение по возрасту будет резко отличаться от общего (рис. 21).

---

[51] Интерпретация и анализ данных в социологических исследованиях. С. 31.

[52] Подробнее об этом см.: Елисеева И.И., Рукавишников О.В. Группировка, корреляция, распознавание образов. М.: Статистика, 1977. С. 82 – 89; Интерпретация и анализ социологических данных. С. 31 – 32; Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. С. 65 – 84; Рабочая книга социолога. М.: Наука, 1983. С. 169 – 172, 190; Статистические методы анализа информации в социологических исследованиях. С. 117 – 120.

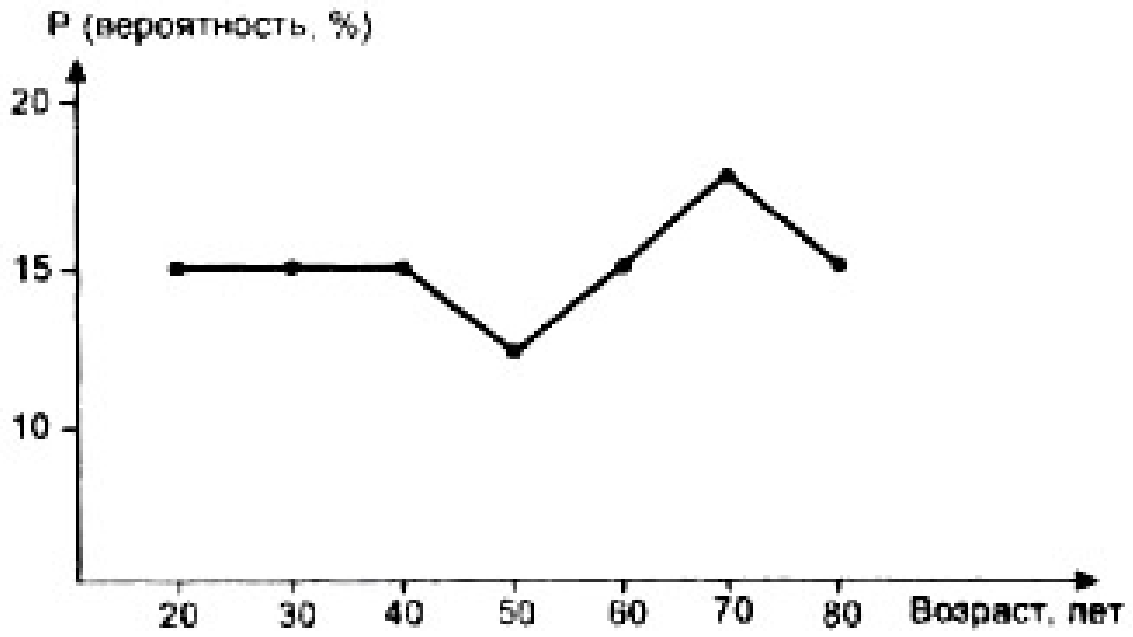


Рис. 20. Гипотетическое распределение по возрасту жителей города N старше 20 лет

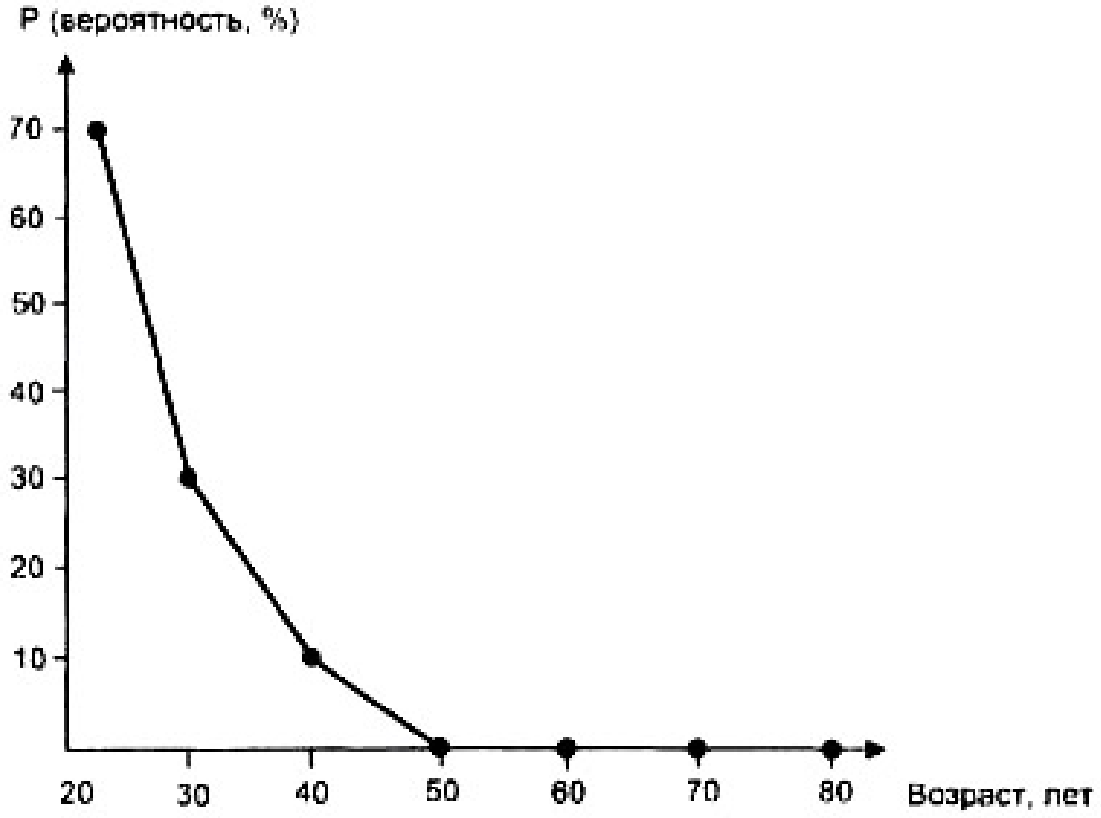


Рис. 21. Гипотетическое распределение по возрасту студентов города N старше 20 лет

Теперь, случайным образом отобрав человека (студента), мы с уверенностью 90% будем полагать, что его возраст не превысит 30 лет.

Мы улучшили возможность прогноза возраста жителей города и можем сделать вывод о наличии связи между признаком «возраст» и признаком «студент». Чтобы сделать этот вывод, мы *сравнили* безусловное распределение признака «возраст» (рис. 20) с его условным распределением (рис. 21), условие состоит в фиксации значения «студент» второго признака.

Все прогнозные 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. Коэффициенты связи, рассматриваемые в данном пособии, отличаются друг от друга способом формализации прогноза.

Выделяют модальный и пропорциональный прогноз.

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 (модальное значение). Такой прогноз называется **модальным (оптимальным).** Коэффициентов для него три: λ_r – отражающий влияние строкового признака на столбцовый; λ_c – отражающий влияние столбцового признака на строковый, λ – усредненный коэффициент.

Рассмотрим формулу для λ_r, (для λ_c рассуждения аналогичны):

$$\lambda_r = \frac{\sum_{i=1}^{r} \max_j n_{ij} - \max_j n_{\cdot j}}{n - \max_j n_{\cdot j}}$$

где выражение $\max_j n_{ij}$ означает наибольшую частоту в i-й строке, из нее мы вычитаем наибольшую столбцовую маргинальную частоту.

Пусть частотная таблица имеет вид:

Таблица 23

Значения признаков X и Y для расчета коэффициента λ_r

Значения X	Значения 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 равен 65. Общее количество объектов в выборке – 150. Таким образом,

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

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 определяем, что упомянутое значение равно 3 (именно это значение является модой). Перебирая последовательно всех респондентов, мы дадим правильный прогноз в 65 случаях и ошибемся в (150 – 65) случаях (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).

Пусть X=1. Соответствующее **условное распределение** Y определяется частотами первой строки матрицы: числами 0, 20, 30. Значит, перебирая 50 респондентов с первым значением X и делая для каждого прогноз, мы не ошибемся в 30 случаях. При X=2 количество верных предположений тоже будет равно 30. При X=3 получим 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. Это – числитель выражения для λ_r. В числителе отражена суть коэффициента, знаменатель же использован для нормировки. Чем ближе значение λ_r к 1, тем лучше предсказание и сильнее связь между переменными. $\lambda_r = 0$, если максимальные частоты в строках приходятся на один столбец. Коэффициенты чаще всего называют коэффициентами Гуттмана[53], Гудмена[54] или λ-коэффициентами[55].

Теперь приведем пример **пропорционального прогноза**[56]. Сначала рассмотрим безусловное распределение**.** 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. Берем случайного респондента, т.е. опускаем руку в урну и вытаскиваем тот шар, который попался случайно. То, что на нем написано, и будет прогнозным значением признака Y для выбранного респондента. Аналогичным образом поступаем и для каждого условного распределения: 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 и шары с единицей, поскольку количество таких шаров равно 0. В соответствии с третьим распределением (X=3) значения 2 и 3 признаков 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. Такие распределения рассматриваются как основа **коэффициента Валлиса**, но принцип его работы тот же, что и у коэффициентов λ.

---

[53] Статистические методы анализа информации в социологических исследованиях. С. 126.

[54] Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. С. 47.

[55] Рабочая книга социолога. 1983.

[56] Толстова Ю.Н. Анализ социологических данных. С. 110.

6.2.3. Коэффициенты связи, основанные на понятии энтропии

В основе этих коэффициентов также лежит сравнение безусловного распределения с условными с точки зрения изучения изменения степени неопределенности рассматриваемых распределений.

По аналогии с энтропией распределения одного признака определяется **энтропия двухмерного распределения:**

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

или

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

Точно так же можно определить энтропию любого многомерного распределения.

Необходимо дать определение еще одного важного понятия – т.н. **условной энтропии:**

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

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

Противоположным понятию энтропии является понятие информации. 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.

Если $H(Y) = 0$, то исход заранее известен. Большее или меньшее значение $H(Y)$ означает большую или меньшую проблематичность результата. Измерение признака X, предшествующее нашему опыту по измерению Y, может уменьшить количество возможных исходов опыта и тем самым снизить степень его неопределенности. Для того чтобы результат измерения X мог сказаться на 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) 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 (термин Шеннона).

Приведенные рассуждения о смысле понятия информации очевидным образом отвечают описанной выше логике сравнения безусловного и условных распределений Y. В основе всех информационных мер связи лежит та разность, которая стоит в правой части последнего равенства. Но именно эта разность и говорит о различии упомянутых распределений. H (Y/X) это обычное среднее взвешенное значение условных энтропий – каждому значению признака X отвечает своя условная энтропия Y:

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

причем каждое слагаемое берется с весом, равным вероятности появления соответствующего условного распределения, т.е. вероятности P_i. Существует ряд мер связи, основанных на понятии энтропии. Например, это I (X, Y) (ненаправленная мера); ее можно интерпретировать как относительное приращение информации об X, возникающее за счет знания Y[57]. Относительность возникает в результате соотнесения такого приращения с первоначальной неопределенностью распределения X. Известны и направленные меры связи:

[57] Миркин Б.Г. Анализ качественных признаков и структур. С. 103.

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

Коэффициенты С называют **асимметричными коэффициентами неопределенности, коэффициентами нормированной информации**[58]**.** $C_{X/Y} = 0$, если и только если переменные X и Y независимы; $C_{X/Y} = 1$, только если X однозначно определяется значением Y (т. е. если полная связь). Аналогичен и коэффициент $C_{Y/X}$.

Соответствующий симметризованный коэффициент нормированной информации вводится следующим образом[59]:

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

Часто используется также **коэффициент Райского**:

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

Он заключен в интервале от 0 до 1; в 0 коэффициент обращается только когда признаки статистически независимы; в 1 – когда признаки полностью детерминируют друг друга.

Информационные меры связи похожи на обычный коэффициент корреляции. Но они имеют одно преимущество: если коэффициент корреляции равен 0, из этого не следует статистическая независимость рассматриваемых признаков; если информационные меры связи равны 0 – из этого следует статистическая независимость рассматриваемых признаков.

6.2.4. Коэффициенты связи для четырехклеточных таблиц сопряженности

Четырехклеточные таблицы – это частотные таблицы, построенные для двух дихотомических признаков, они представля-

[58] Елисеева И.И., Рукавишников О.В. Группировка, корреляция, распознавание образов. С. 91.

[59] Там же. С. 95.

ют собой частный случай таблиц сопряженности. Пусть рассматриваются два дихотомических признака – пол (1 – мужчина, 0 – женщина) и курение (1 – курит, 0 – не курит). Буквы в клетках таблиц обозначают соответствующие частоты (см. табл. 24, 25).

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**. Равенство $ad = bc$ эквивалентно равенству $\frac{a}{c} = \frac{b}{d}$, что, в свою очередь, означает пропорциональность столбцов (строк) частотной таблицы, т. е. отсутствие статистической связи. Можно показать, что разница между наблюдаемой и теоретической частотой для левой верхней клетки нашей четырехклеточной частотной таблицы (наличие или отсутствие связи для такой таблицы определяется содержанием единственной клетки – при заданных маргиналах частоты, стоящие в других клетках, можно определить однозначно) равна величине[60]

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

Таблица 24

Общий вид четырехклеточной таблицы сопряженности

Значения X	Значения Y		Итого
	1	2	
1	a	b	1
0	c	d	0
Итого	a+c	b+d	Итого

[60] Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973. С. 722.

Таблица 25

Конкретизированный вид четырехклеточной таблицы сопряженности

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

Коэффициенты всегда 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 0, во втором – близость отношения к 1. В обоих случаях требуется нормировка. И желательно, чтобы искомые показатели связи находились либо в интервале от –1 до 1, либо от 0 до 1. Есть разные коэффициенты связи:

Коэффициент ассоциации Юла вычисляется как

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

Коэффициент контингенции вычисляется как

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

Оба коэффициента изменяются в интервале от –1 до +1 (определяем направленность связи), обращаются в 0 в случае отсутствия статистической зависимости, в 1 или –1 эти коэффициенты обращаются в разных ситуациях. Они схематично отражены ниже (табл. 26).

Та связь, которую отражает Q, названа *полной,* которую отражает Ф – *абсолютной.* Иногда используют иную терминологию: говорят, что Q измеряет одностороннюю связь, Ф – двустороннюю.

Таблица 26

Схематическое изображение свойств коэффициентов Q и Ф

Свойства коэффициентов	Q = 1		Q = –1		Ф = 1		Ф = –1	
Отвечающие им виды таблиц	a	0	0	b	a	0	0	b
	c	d	c	d	0	d	c	0
	a	b	a	b				
	0	d	c	0				

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 = 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. Если же все A являются одновременно B и все B являются одновременно A, то связь называется абсолютной.

Поясним смысл знака рассматриваемой связи. «Положительность» означает, что какое-то значение первого признака сопрягается с одним значением другого, а «отрицательность» – с другим (при наличии положительной связи все мужчины курят, а 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 мы имеем дело с данными, представленными табл. 27:

Таблица 27

Частотная таблица для демонстрации отношения преобладаний

Курение	Пол		Итого
	м	ж	
Курит	50	90	140
Не курит	20	40	60
Итого	70	130	200

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 больше курящих (50:20), 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** [61]:

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

или в общем случае $\lambda = \frac{a:c}{b:d}$.

Если отношение преобладания больше единицы, то связь называется *положительной,* если меньше единицы – *отрицательной.*

Если мы, используя обозначения 0 и 1 для значений наших признаков, будем интерпретировать эти обозначения как настоящие числа, то обычный коэффициент корреляции между признаками окажется равным Ф. Этот факт имеет огромное значение для анализа данных. Одним из популярных способов создания возможности использования числовых математико-статистических методов для анализа номинальных данных является т.н. дихотомизация: замена (по определенным правилам) одного номинального признака таким количеством дихотомических, принимающих значения 0 и 1, сколько в нем альтернатив и дальнейшая работа с 0 и 1 как с обычными числами[62].

За каждым коэффициентом стоит своя модель, свое понимание этой связи. И для того, чтобы найти связь, надо использовать целый набор коэффициентов.

---

[61] Rudas T. Odds ratios in the analysis of contingency tables. Sage university paper series on Quantitative applications in the social sciences, 07-119. Newbury park, CA: Sage, 1998.

[62] Интерпретация и анализ данных в социологических исследованиях. С. 29 – 30; Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. С. 84 – 93; Рабочая книга социолога. С. 189; Статистические методы анализа информации в социологических исследованиях. С. 116 – 117.

6.2.5. Многомерные отношения преобладаний

В реальности двухмерных связей практически не существует: они многомерны. 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 значение четвертой переменной и т.д.

В работе Б.Г. Миркина[63] приводится пример того, как при фиксации значения третьей переменной обусловливается «возникновение» связи между двумя переменными. Изучалась зависимость между наличием в семьях пылесоса (П) и холодильника (Х). Исходная частотная таблица имела вид:

	П	мП	Маргинал
Х	560	840	1 400
мХ	240	360	600
Маргинал	800	1 200	2 000

Зависимость здесь 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, т.е. семей с высоким (Д) и низким (мД) уровнем дохода. Получились следующие две частотные таблицы.

Для Д:

	П	мП	Маргинал
Х	520	300	820
мХ	80	100	180
Маргинал	600	400	1 000

[63] Миркин Б.Г. Группировки в социально-экономических исследованиях. С. 18 – 20.

Для мД:

	П	мП	Маргинал
Х	40	540	580
мХ	160	260	420
Маргинал	200	800	1 000

В обоих случаях связь присутствует (пропорциональности строк нет). Более того, для первой таблицы она положительна (значение Х сопрягается со значением П: семьи, имеющие холодильник, как правило, имеют и пылесос), а для второй – отрицательна (значение Х сопрягается со значением мП: семьи, имеющие холодильник, чаще всего не могут купить пылесос).

В таблице, отвечающей высокому доходу Д, отношение преобладания $\frac{520:80}{300:100}=\frac{13}{6}$, т.е. больше единицы, а в таблице, отвечающей низкому доходу, аналогичное отношение $\frac{40:160}{540:260}=\frac{13}{108}$, т.е. меньше единицы.

В работе Г. Аптона[64] приводится пример т.н. парадокса Симпсона. Исходная таблица имела вид

	В	мВ	Маргинал
А	495	805	1300
мА	405	295	700
Маргинал	900	1 100	2 000

В ней наблюдается явная отрицательная связь: отношение преобладаний $\frac{495:405}{805:295}=0,45$ меньше единицы (значение А имеет большую тенденцию встречаться с мВ, 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.

[64] Цит. по: Толстова Ю.Н. Анализ социологических данных. С. 74.

Для С:

	В	мВ	Маргинал
А	95	800	895
мА	5	100	105
Маргинал	100	900	1 000

Для мС:

	В	мВ	Маргинал
А	400	5	405
мА	400	195	595
Маргинал	800	200	1 000

Соответствующие отношения преобладаний равны:

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

На основании рассмотренных выше примеров выделим три модели работы с дихотомическими признаками.

1. Если 1 дихотомический признак.

P_1 – доля объектов, обладающих первым значением признака, P_2 – вторым. Соответствующее отношение преобладания первого порядка

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

будет обозначать, во сколько раз объем первого множества больше (меньше) второго. Если отношение преобладания больше 1, мы имеем дело с положительным преобладанием, если меньше – с отрицательным.

2. Если 2 дихотомических признака.

P_{11} – доля объектов с первым значением первого признака и первым значением второго, P_{12} – с первым значением первого и вторым значением второго и т.д. Двухмерная частотная таблица приобретет вид:

$$P_{11} \qquad P_{12}$$
$$P_{21} \qquad P_{22}$$

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:

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

Проверяем, в какой мере столбцы таблицы сопряженности являются пропорциональными. Если λ_2 равно единице, то двухмерной связи нет. Если больше единицы, то говорят о положительной связи (и чем больше отличие от 1, тем больше эта связь). Если λ_2 меньше 1, то говорят об отрицательной связи.

λ_2 – это отношение двух λ_1 для первого признака, вычисленных отдельно для каждого из двух значений второго признака.

3. Если 3 дихотомических признака.

Фиксируем первое значение третьего признака и вычисляем λ_2 по первым двум признакам:

$$\frac{P_{111} : P_{211}}{P_{122} : P_{222}}.$$

Аналогичную величину вычисляем, фиксируя второе значение третьего признака:

$$\frac{P_{112} : P_{212}}{P_{122} : P_{222}}.$$

Находим отношение последних двух величин. Это – отношение преобладания третьего порядка:

$$\lambda_2 = \frac{\dfrac{P_{111} : P_{211}}{P_{121} : P_{221}}}{\dfrac{P_{112} : P_{212}}{P_{122} : P_{222}}}.$$

Если отношения преобладания второго порядка, вычисленные для каждого из двух значений третьего признака, были примерно одинаковыми, то λ_3 будет примерно равно 1. Это означает отсутствие трехмерной связи. Если λ_3 больше 1, говорят о положительной трехмерной связи; если λ_3 меньше – об отрицательной трехмерной связи и т.д.

6.3. Анализ связей типа «альтернатива – альтернатива»: ДА

Для изучения такой связи мы вводим понятие **локальной связи.** Это связь между отдельными альтернативами рассматриваемых признаков. Локальному подходу отвечает понимание связи как некоторого отношения между двумя конкретными градациями a и b признаков X и Y соответственно. Связь сильная, если из того, что для некоторого объекта первый признак принимает значение a, с большой вероятностью следует, что второй признак для того же объекта принимает значение b. Если вероятность мала – она слаба.

Рассмотрим частотную таблицу, выражающую зависимость между профессией человека и читаемой им газетой; для простоты предполагаем, что каждый респондент может читать не более одной газеты (табл. 28).

Таблица 28

Связь между профессией респондента и выбором им газеты

Профессия	Читаемая газета				Итого
	УГ	МК	Независимая	Правда	
Врач	5	2	13	8	28
Токарь	6	24	7	13	50
Учитель	9	0	1	0	10
Космонавт	2	1	4	5	12
Итого	22	27	25	26	100

Нас интересует локальная связь между свойством «быть учителем» и свойством «читать "Учительскую газету" (УГ)». Упомянутая выше четырехклеточная таблица будет иметь вид:

Таблица 29

Связь между свойством «быть учителем» и свойством «читать УГ»

Профессия	Читаемая газета		Маргиналы по строкам
	УГ	не УГ	
Учитель	9	1	10
Не учитель	13	77	90
Маргиналы по столбцам	22	78	100

При дальнейшем использовании коэффициентов мы и измерим силу локальной связи. Для этого используют понятие детерминационного анализа (**ДА-алгоритм**), которое лежит в основе множества статпакетов и широко применяется в прикладной социологии.

Детерминация $a \to b$ характеризуется интенсивностью (точностью, истинностью) $I(a \to b) = P(b/a)$ и емкостью (полнотой) $C(a \to b) = P(a/b)$.

$$I(a \to b) = P(b/a) = P(УГ/учитель) = \frac{9}{10} = 0,9.$$

Это у нас доля читающих УГ среди учителей (90%).

$$C(a \to b) = P(a/b) = P(УГ/учитель) = \frac{9}{22} \approx 0,41.$$

Это у нас доля учителей среди читающих УГ (41%).

Вычисление интенсивности и емкости изучаемых детерминаций – основной элемент детерминационного анализа.

В качестве объясняющего признака могут выступать конъюнкции и дизъюнкции любых значений рассматриваемых признаков-предикторов.

«Точность правила *Если a, то b* вычисляется по формуле[65]:

[65] ДА-система (Детерминационный анализ). М.: Контекст. 1997. С. 160 – 167.

$$\frac{N(a,b)}{N(a)}.$$

где N (a, b) – количество объектов, обладающих одновременно объясняющим признаком a и объясняемым признаком b (количество подтверждений правила); N(a) – количество объектов, обладающих объясняющим признаком a безотносительно к любым другим признакам (количество применений правила). Точность измеряется от 0 до 1. Точность правила *Если a, то b* есть мера достаточности a для наличия b. Точность правила – это главный критерий его практической ценности. Наиболее ценятся правила, имеющие точность, близкую к 1.

Полнота правила вычисляется по формуле:

$$\frac{N(a,b)}{N(b)},$$

где N (b) количество объектов, обладающих объясняемым признаком b безотносительно к любым другим признакам (объем объясняемого признака). Полнота изменяется от 0 до 1. Полнота правила *Если a, то b* есть мера необходимости a для наличия b. Полнота правила – это второй по значимости (после точности) критерий его практической ценности. Предельно точные правила ценятся тем выше, чем больше их полнота. Однако наличие высокой полноты не обязательно. Система точных правил, каждое из которых имеет небольшую полноту, может иметь чрезвычайную полезность для практики и науки, если ее суммарная полнота близка к 1».

Приведем пример, где объясняемое положение – голосование за кандидата N. Допустим, что 40% мужчин проголосовали за N. Это значит, что точность правила «если мужчина, то голосует за N» равна 0,4. Если мы рассмотрим мужчин с высшим образованием, точность детерминации может повыситься: за N проголосовали 80% мужчин с высшим образованием.

Если какой-либо объясняющий признак убрать из правила, точность правила изменится. Величина этого изменения (с учетом знака) и есть вклад объясняющего признака в точность. Рассмотрим правило *Если a и b, то c.* Вклад S (a) объясняющего признака в точность правила вычисляется по формуле:

S(*a*) = (Точность правила *Если a и b, то c*) –
(Точность правила *Если b, то c*).

Аналогично вычисляется вклад любого объясняющего признака в точность в любом заданном правиле. Аналогично определяется вклад Q(a) объясняющего признака в полноту правила[66].

6.4. Анализ связей типа «группа альтернатив – группа альтернатив»

Такие связи социологу необходимо устанавливать при анализе групп. Например, нужно проанализировать зависимость между свойствами «быть учителем, или врачом, или научным сотрудником» и «читать ЛГ или журнал Новый Мир».

Проблемы возникают, если мы не фиксируем заранее указанную подтаблицу, а ставим перед собой цель, например, найти такие подтаблицы исходной таблицы сопряженности, которые обладают свойствами, отличающими их от всей таблицы (либо от других подтаблиц).

Мы рассмотрим 2 основных направления применения математики для этих целей – анализ фрагментов таблицы и методы поиска сочетания независимых предикторов.

6.4.1. Анализ фрагментов таблиц сопряженности

Существует возможность такого разложения исходной частотной таблицы на четырехклеточные подтаблицы, что исходный хи-квадрат будет приблизительно равен сумме «четырехклеточных» хи-квадратов. При этом количество упомянутых подтаблиц равно числу степеней свободы исходной таблицы:

$$\chi^2 \approx \sum_i \chi_i^2 ,$$

где χ_i отвечает i-й четырехклеточной компонентной подтаблице (т.е. подтаблице). При расчете хи-квадрат мы как бы суммируем,

[66] ДА-система (Детерминационный анализ). С. 160 – 167.

усредняем отдельные «клеточные» отклонения. Соотношение говорит о том, какой именно вклад в общее отклонение частот от условия статистической независимости дают фрагменты такого рода.

Разложение $\chi^2 \approx \sum_i \chi_i^2$ ничего не даст социологу, если все «четырехклеточные» хи-квадраты превышают (или все не превышают) соответствующие табличные критические значения, т.е. если для всех наших компонентных подтаблиц мы должны отвергнуть (или для всех же принять) нуль-гипотезу о независимости соответствующих пар альтернатив друг от друга. Тогда и исходный хи-квадрат превышает (не превышает) отвечающее ему табличное значение, и мы можем считать, что отвержение (принятие) нуль-гипотезы как бы равномерно опирается на все значения рассматриваемых признаков. Считаем, что в таком случае никаких интересных подсвязей исходная таблица не содержит.

Другое дело, если одни «четырехклеточные» хи-квадраты будут превышать соответствующие критические значения, а другие – не будут (из десяти подтаблиц только для трех имеются основания отвергнуть отвечающую им нуль-гипотезу, значит, исходный хи-квадрат отличается от нуля (показывает отклонение ситуации от состояния статистической независимости признаков) за счет наличия связи именно в этих трех подтаблицах, остальные же подтаблицы к наличию связи не имеют отношения).

Подтаблица может получаться за счет вырезания соответствующего фрагмента из исходной матрицы сопряженности или в результате суммирования определенных строк и столбцов. Мы получали из исходной таблицы четырехклеточную таблицу сопряженности – в клетке, отвечающей сочетанию «не учитель, читает УГ» стояла частота, полученная из исходной таблицы путем суммирования всех респондентов, читающих УГ, но имеющих профессии, отличные от профессии учителя и т.д. (табл. 30).

Таблица 30

Схематическое изображение четырехклеточного фрагмента таблицы

	Читает УГ	Не читает УГ
Учитель	Исходная частота	Сумма респондентов-учителей, читающих газеты, отличные от УГ
Не учитель	Сумма респондентов, являющихся не учителями и читающих УГ	Сумма респондентов, являющихся не учителями и читающих газеты, отличные от УГ

Правила получения компонентных четырехклеточных фрагментов таковы[67]:

1. Каждая из частот исходной таблицы должна встречаться только в одной из компонентных таблиц.

2. Маргинальные частоты исходной таблицы должны встречаться в одной из компонентных таблиц как частоты определенного типа: либо как стоящие в клетке частотной таблицы, либо как маргинальные.

3. Каждая частота, содержащаяся в одной из компонентных таблиц, но отсутствующая в исходной таблице, должна появиться в другой компонентной таблице как частота другого типа – «клеточная», если была маргинальной, и наоборот.

Какое из возможных разложений мы выберем для интерпретации, определяется задачей исследования.

По данным обследования семейных групп (семья сына или дочери – семья родителей)[68] рассмотрим зависимость характера желаемого расселения (отделения молодой семьи от семьи родителей) от состава молодой семьи и возраста женщины в этой семье (табл. 31).

---

[67] Интерпретация и анализ данных в социологических исследованиях. С. 43 – 44.

[68] Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. С. 86.

Таблица 31

Зависимость характера желаемого расселения от состава молодой семьи и возраста женщины

Характеристика молодой семьи		Желаемое расселение			Итого
возраст женщины, лет	состав	в одной квартире	в разных квартирах	в одном микрорайоне и дальше	
До 30	Мать с детьми	6	8	6	20
	Брачная пара с детьми	11	112	66	189
30-40	Мать с детьми	6	12	18	36
	Брачная пара с детьми	24	122	121	267
40-55	Мать с детьми	5	5	8	18
	Брачная пара с детьми	8	23	8	39
Итого		60	282	227	569

χ^2=39,2 в то время, как $\chi^2_{табл}$=18,3 (a = 0,05; df = 10). Отвергаем нуль-гипотезу. Но, возможно, существует связь между некоторыми наборами альтернатив. Построим разложение исходной таблицы на четырехклеточные:

Разложение таблицы 31 на подтаблицы

6	14	20
54	495	549
60	509	569

(А)

11	178	189
43	317	360
54	495	549

(В)

8	6	14
274	221	495
282	227	509

(Б)

112	66	178
162	155	317
274	221	495

(Г)

6	30	36
37	287	324
43	317	360

(Д)

122	121	243
28	16	44
150	137	287

(З)

12	18	30
150	137	287
162	155	317

(Е)

5	13	18
8	31	39
13	44	57

(И)

24	243	267
13	44	57
37	287	324

(Ж)

5	8	13
23	8	31
28	16	44

(К)

Лишь для 5 из 10 получившихся четырехклеточных таблиц соответствующее значение χ^2 превышает табличное, отвечающее 5%-му уровню значимости (это значение будет отличаться от приведенного выше из-за различия числа соответствующих степеней свободы: для исходной таблицы это число равно 10, а для четырехклеточной – равно в данном случае $\chi^2_{табл}$ = 3,8. Частоты, отвечающие значению первого признака «остальные» из таблицы (А), получаются путем суммирования строк исходной таблицы, соответствующих всем рассматриваемым сочетаниям значений двух наших характеристик молодой семьи, кроме сочетания «женщина с детьми, до 30 лет»; частоты, отвечающие значению второго признака «в разных квартирах», получаются за счет суммирования столбцов исходной матрицы «в одном доме» и «в одном микрорайоне и дальше» и т.д. Критический уровень превышают критерии χ^2, отвечающие таблицам (А), (В), (Г), (Ж), (К). Сумма этих критериев равна 33, 9, что хотя и не равно значению χ^2 для исходной таблицы (39, 2), но составляет от него почти 86% (отклонение эмпирических частот от теоретических в исходной таблице на 86% объясняется наличием связи).

Таблица 32

Сочетания признаков (компонентные подтаблицы)

1-й признак	2-й признак	Обозначение подтаблицы
Мать с детьми, до 30 лет	в одной квартире	(А)
Остальные	в разных квартирах	
То же	в одном доме	(Б)
	дальше	
Брачная пара, мать до 30 лет	в одной квартире	(В)
Остальные	в разных квартирах	
То же	в одном доме	(Г)
	дальше	
Мать с детьми, 30 – 40 лет	в одной квартире	(Д)
Остальные	в разных квартирах	
То же	в одном доме	(Е)
	дальше	
Брачная пара, мать 30 – 40 лет	в одной квартире	(Ж)
Остальные	в разных квартирах	
То же	в одном доме	(З)
	дальше	
Мать с детьми, 40 – 55 лет	в одной квартире	(И)
Брачная пара, 40 – 55 лет	в разных квартирах	
То же	в одном доме	(К)
	дальше	

Рассмотрим подтаблицу (А):

(А)

Тип молодой семьи	Желаемое расселение		Итого
	в одной квартире	в разных квартирах	
Мать с детьми, до 30 лет	6	14	20
Остальные	54	495	549
Итого	60	509	569

Значение χ^2 для этой подтаблицы равно 8,3, что превышает табличное значение, равное 3,8. Отступление от ситуации незави-

симости происходит за счет того, что доля желающих остаться в одной квартире со старшим поколением молодых матерей-одиночек (таких молодых матерей-одиночек почти треть: 6 из 20) выше, чем аналогичная доля среди всех опрошенных (среди всех опрошенных не хотят разъезжаться с бабушками-дедушками лишь чуть более 10%: 60 из 569). Вывод: для семей, состоящих из молодых матерей-одиночек с детьми, вопрос о необходимости разъезжаться со старшим поколением стоит менее остро, чем для других категорий семей.

Для подтаблиц (Б) и (Д) χ^2 (равные соответственно 0,02 и 0,8) и не превышают критических значений:

(Б)

Тип молодой семьи	Желаемое расселение		Итого
	в одном доме	дальше	
Мать с детьми, до 30 лет	8	6	14
Остальные	274	221	495
Итого	282	227	509

(Д)

Тип молодой семьи	Желаемое расселение		Итого
	в одном доме	дальше	
Мать с детьми, до 30 лет	6	30	36
Остальные	37	287	324
Итого	43	317	360

Подтаблица (Б) говорит о том, что молодые матери-одиночки примерно в той же мере выбирают те или иные варианты расселения, что и семьи других типов. Другими словами, соответствующая специфика семьи не сказывается в том, хочет ли желающая переселиться молодая семья после переезда остаться поближе к родителям (в одном доме) или же готова уехать подальше. Чуть более половины желающих разъехаться хотят остаться в одном доме со старшими (282 из 509), так же как и среди матерей-одиночек до 30 лет (8 из 14).

При анализе подтаблицы (Д) ясно, что для более старших матерей одиночек (30 – 40 лет) указанной выше специфики в жела-

нии расселиться нет: семьи этой категории ровно в той же мере хотят разъезда (6 из 36 семей не хотят отделяться от старших), как и семьи других типов (не хотят разъезжаться 37 из 324).

6.4.2. Методы поиска сочетаний значений независимых признаков (предикторов)

Допустим, перед нами огромный массив информации, скажем 1 000 заполненных анкет по 30 вопросов в каждой. При изучении причинно-следственных отношений естественно выделение, с одной стороны, некоторых признаков Y, которые описывают основное интересующее исследователя явление, а с другой – совокупности признаков X, потенциально являющихся причинами. Y – зависимые переменные (объясняемые, детерминируемые, целевые, критериальные, результирующие, признаки-следствия, функции). Социолога интересует, какими факторами (причинами) определяется некоторое поведение респондента. Y может состоять в том, что респондент в ответе на один из вопросов анкеты выражает свою готовность проголосовать на выборах за кандидата Ж. **Мы должны установить, какими сочетаниями значений рассматриваемых признаков обладают эти люди** (исследователь должен перебрать все возможные сочетания значений рассматриваемых признаков и найти среди них такие, обладателям которых присуще рассматриваемое поведение). Однако в действительности это тяжело сделать по причине огромных временных затрат (мы не знаем какие признаки взять, сколько их, какие сочетания значений каждого признака следует принять во внимание, а группу, где 100% людей обладает интересующим нас свойством, мы не найдем из-за ненадежности нашего способа измерения мнений респондентов (анкетный опрос)). Будем называть ту или иную группу респондентов типом, олицетворяющим интересующее нас поведение, если для этой группы удовлетворяется выбранный нами критерий (более высокое качество будет иметь та группа, где доля желающих голосовать за Ж выше).

После этого перебираем всевозможные сочетания значений рассматриваемых признаков и для каждого из них проверяем, можно ли соответствующую совокупность объектов считать оли-

цетворением определенного типа поведения. Если нет – переходим к проверке следующего сочетания значений, если да – считаем, что нашли группу-тип.

Кроме того, неясно, при каких условиях считать, что мы нашли группу, обладающую указанным поведением: если среди этих людей желают проголосовать за Ж 90 или 85%[69]?

Существуют разные способы решения задачи. Западные авторы предложили группу алгоритмов, в названия которых входит аббревиатура AID (automatic interaction detector). Наличие сравнительно большого количества таких алгоритмов объясняется тем, что решение аналогичных задач очень актуально для прикладных социологических исследований. Рассмотрим два алгоритма.

Алгоритм THAID. Допустим, задан некоторый номинальный признак Y «За кого Вы собираетесь голосовать?» с пятью альтернативами-вариантами ответов: Е, Ж, З, Л, Я. Для каждой проверяемой группы объектов будем вычислять распределение входящих в нее респондентов по этому признаку, подсчитывать соответствующее модальное значение и определять долю его встречаемости. Соответствующий процент будет служить оценкой качества группы как типа. Пусть распределения в 2 группах следующие (рис. 22):

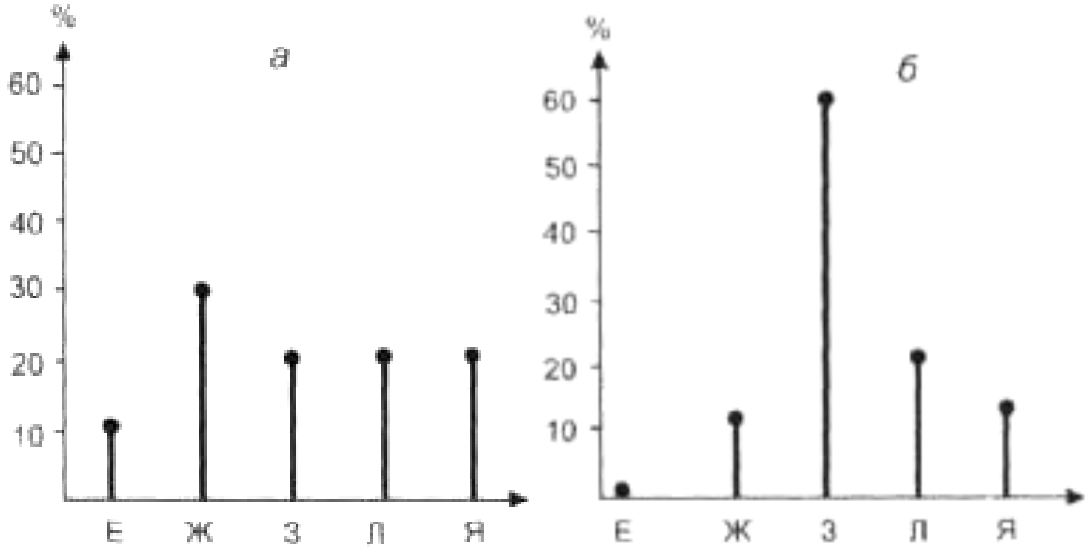


Рис. 22. Частотные распределения, отражающие электоральное поведение двух групп респондентов

[69] Толстова Ю.Н. Анализ социологических данных. С. 152.

Модальное значение для первой совокупности – Ж, его доля – 30 %. Для второй совокупности мода – З. Ее доля – 60%. Качество второй совокупности выше. Однако, вероятно, мы ни ту, ни другую группу не можем рассматривать как тип (процент низок). Работаем с каждым признаком отдельно.

Сначала перебираем варианты разбиения всех альтернатив признака на две части: (первая – все остальные); (первая и вторая – все остальные); (первая, вторая, третья – все остальные) и т.д. до последнего варианта: (все, кроме последней – последняя). Множество значения разбивается только на *две* части и «склеиваются» только *соседние* градации.

Оцениваем качество как долю модальной частоты признака-функции каждой из двух групп, получающихся при одном разбиении одного признака (имеются ввиду группы респондентов, отметивших альтернативы той или иной группы; мы как бы отождествляем группу альтернатив и группу отвечающих им респондентов). Пусть первая группа включает n_1 человек и доля модальной частоты для нее составляет P_1 %, а вторая группа состоит из n_2 человек и доля модальной частоты составляет P_2 %. Тогда вычислим показатель качества всего разбиения:

$$P = n_1 \times P_1 + n_2 \times P_2 .$$

Мы имеем дело со взвешенным средним (такой способ усреднения очень распространен в социологии). Из каждого разбиения совокупности альтернатив каждого признака выберем лучшее. Скажем, таковым оказалось разбиение совокупности альтернатив признака «образование» на группы (1, 2) и (3, 4, 5). Далее будем изучать респондентов каждой группы отдельно.

Берем респондентов с низким образованием (1, 2) и делаем для них то же самое. Получим самое хорошее разбиение совокупности респондентов – скажем, это будет разбиение по признаку «семейное положение», группы альтернатив (1, 2) и (3). Далее должны рассмотреть людей с высоким образованием (отметивших альтернативы 3, 4, 5 – среднее, неполное высшее и высшее образование соответственно) и реализовать для них ту же процедуру. Допустим, для них наилучшим оказалось разбиение по социальному происхождению группы альтернатив (1) и (2 и 3).

Будем изучать отдельно тех людей с низким образованием, которые женаты или неженаты (альтернативы 1 и 2), и тех людей с низким образованием, которые разведены (альтернатива 3). Отдельно исследуем группы людей с высоким образованием из семей рабочих (альтернатива 1) и людей с высоким образованием, из семей служащих или военных (альтернативы 2, 3). Каждая из четырех получившихся групп разделится еще на две. И каждый раз мы будем получать группы с увеличивающейся долей модальной частоты по нашему признаку. Остановиться мы можем в следующих случаях:

а) найдена группа с большой долей модальной частоты (среди людей с низким образованием и разведенных 95% проголосовали за Л – следовательно, тип найден);

б) получилась слишком малочисленная группа – игнорируем это и двигаемся дальше, исключив соответствующих людей из рассмотрения, или выясняем, в чем состоят особенности этих людей, изучаем их;

в) получилась слишком длинная цепочка – вряд ли мы сделаем серьезные выводы на основе знания того факта, что люди с высоким образованием, неженатые, живущие в сельской местности, имеющие более 4-х детей, 3-х поросят, не любящие смотреть телевизор и мечтающие о путешествии на Кипр почти все проголосовали за Л.; по той же причине мы обычно не воспринимаем как закономерность классификацию, в которой 1 500 классов);

г) ПК не нашел ни одной совокупности с интересующими нас свойствами (в анкете не заложено описание этого поведения – такая ситуация может быть следствием нашего неумения составлять анкету, общаться с респондентом, учитывать цели исследования при формировании инструментария, ставить задачу и т.д.)[70].

Алгоритм CHAID. Заданы те же исходные данные, что и при работе алгоритма THAID, задается номинальный признак-функция Y. Но групповое поведение будем ассоциировать не с частотой модального значения признака Y, а со всем распределением этого признака. Как и выше, в нашу задачу наряду с поиском со-

[70] См. подробнее: Интерпретация и анализ данных в социологических исследованиях. С. 29, 136 – 151; Рабочая книга социолога. С. 193 – 195; Типология и классификация в социологических исследованиях. С. 213 – 230.

четаний значений рассматриваемых признаков, детерминирующих групповое поведение, входит поиск конкретных видов такого поведения – конкретных распределений значений признака Y.

Пусть Y – электоральное поведение респондента, а признак X – это профессия с градациями «врач», «учитель», «рабочий». Для определения «склеиваемых» градаций признака «профессия» используем алгоритм CHAID – рассмотрим частотную таблицу, связывающую эти два признака (табл. 33)[71].

Таблица 33

Определение склеиваемых градаций признака «профессия» при голосовании

Профессия	Предполагаемое голосование					Итого
	Е	Ж	З	Л	Я	
Врач	10	2	10	8	30	60
Учитель	5	1	5	4	15	30
Рабочий	0	30	8	20	2	60
Итого	15	33	23	32	47	150

Мы должны склеить следующие градации: респонденты, отметившие одну градацию, обладают тем же поведением, что и респонденты, отметившие другую. Рассмотрение соответствующих совокупностей респондентов отдельно не имеет смысла. Такими свойствами обладают градации «врач» и «учитель». Если мы рассмотрим отдельно представителей этих профессий, то не получим разные типы избирателей: половина врачей хочет голосовать за Я, половина учителей также выбирают Я. Одинаковое количество учителей (5 человек, 17 %) хотят голосовать за Е и З соответственно, и то же самое можно сказать о врачах и т.д. Врачей же и рабочих нельзя объединять. Они являют собой совершенно разный тип электорального поведения: за Я собираются голосовать 50% (30 человек) врачей и менее 2% (2 человека) рабочих и т.д.

Для конкретного признака X проверяем все пары альтернатив. Считаем, что каждая пара отвечает своему дихотомическому признаку и, задавшись уровнем значимости (скажем, а = 0,05),

[71] Толстова Ю.Н. Анализ социологических данных. С. 155 – 157.

вычисляем критерий хи-квадрат для этого признака и Y. Отбираем те пары, для которых значение χ^2 не превышает соответствующее критическое значение. Это пары, для которых имеет смысл принять нашу нуль-гипотезу. Далее выбираем ту пару, для которой χ^2 меньше всего, т. е. для которой наша нуль-гипотеза принимается как бы с большей надежностью. Именно альтернативы этой пары мы и склеиваем.

Склеив какие-то альтернативы в каждом из анализируемых признаков, мы вычисляем критерий хи-квадрат между каждым из оставшихся к рассматриваемому шагу признаком X_i и Y. Отберем те признаки X_i, для которых наш критерий превышает критическое значение, т. е. для которых следует считать, что между каждым из них и Y есть связь. Среди этих признаков отберем тот, для которого χ^2 имеет наибольшее значение (связь существует с наибольшей вероятностью). По его градациям мы и будем далее разбивать совокупность респондентов.

Описанные процедуры мы реализуем так же, как и в алгоритме THAID. В итоге выделяются группы респондентов, каждая из которых описывается последовательностью значений рассматриваемых признаков.

CHAID, так же как и THAID, не гарантирует выявления в исходных данных всех интересующих исследователя закономерностей, т. к. на каждом шаге разбиения алгоритм оценивает лишь двухмерную связь. Алгоритм задействован в известном пакете программ SPSS и очень информативен для социолога.

6.5. Анализ связей типа «признак – группа признаков»

6.5.1. Номинальный регрессионный анализ (НРА)

Иногда имеет смысл искать сочетания значений исходных признаков, которые определяют те или иные связи, то или иное поведение респондентов, или объединять отдельные признаки друг с другом, искать такие их сочетания, которые детерминируют другие признаки. Это позволяет сделать регрессионный анализ.

Пусть нас интересует зависимость между X и Y. Но, зная коэффициент их корреляции, мы не можем сказать, как возрастет значение Y, если значение X увеличится, скажем, на 1.

В качестве примера рассмотрим зависимость между производственным стажем человека и его зарплатой (рис. 23 а и б)[72].

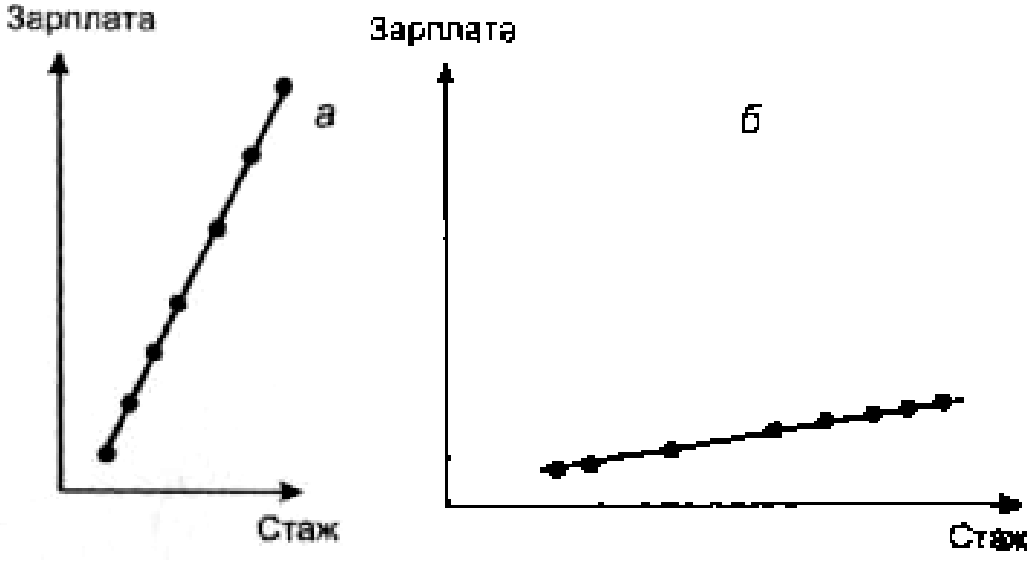


Рис. 23. Сильные линейные связи признаков «зарплата» и «стаж», определяющих разный прогноз

В обоих случаях соответствующие коэффициенты корреляции близки к 1 (обе совокупности точек-объектов лежат на прямых линиях, отвечающих нашей зависимости). На рис. 23а из них прямая идет резко вверх. Поэтому даже при небольшом увеличении X признак Y резко возрастет. В случае же наличия связи, изображенной на рис. 23б, прямая близка к горизонтали. Поэтому даже при значительном росте X значение Y почти не изменится. Это нельзя узнать лишь на основе вычисления коэффициентов корреляции.

Чтобы делать прогноз, как изменится значение Y при том или ином изменении значения X, нам желательно знать форму связи между этими переменными, т. е. функцию вида Y = f (X). Независимые переменные называют *входными, экзогенными, внешними*, а зависимые – *выходными, эндогенными, внутренними*. Если переменные X и Y – независимая и зависимая, то ищем усредненную зависимость вида Y = f (X).

[72] Толстова Ю.Н. Указ. соч. С. 161.

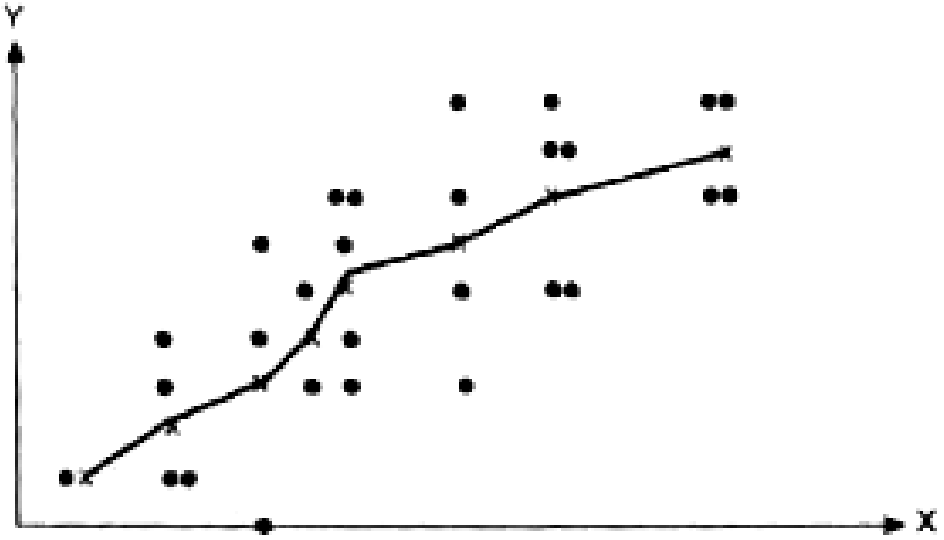


Рис. 24. Принципиальная схема линии регрессии

Для социологических данных типична ситуация, когда одному значению X соответствует множество значений Y. Эта ситуация схематично изображена на рис. 24. Чтобы выбрать четкую зависимость, подсчитаем для каждого значения X среднее арифметическое значение всех отвечающих ему значений Y и будем изучать зависимость от X таких средних. Соответствующие точки на нашем рисунке обозначены крестиками и по ним проведена кривая:

$$\overline{Y}x = f(X).$$

На рисунке отражена выборочная ситуация, в то время как в действительности нас интересует то, что делается в генеральной совокупности. Рассмотрение последней предполагает, что переменные непрерывны, т.е. имеют бесконечное число значений. Соотношение для генеральной совокупности имеет следующий вид:

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

где μ – знак математического ожидания меры средней тенденции для генеральной совокупности. Такая функция называется **функцией регрессии Y по X** (уравнением регрессии, либо регрессионной зависимостью).

Фиксируя какое-либо значение X, равное, например, X_i (рассматривая некую совокупность объектов), мы имеем дело с неко-

торым 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. Математическое ожидание лежит на линии регрессии (рис. 25).

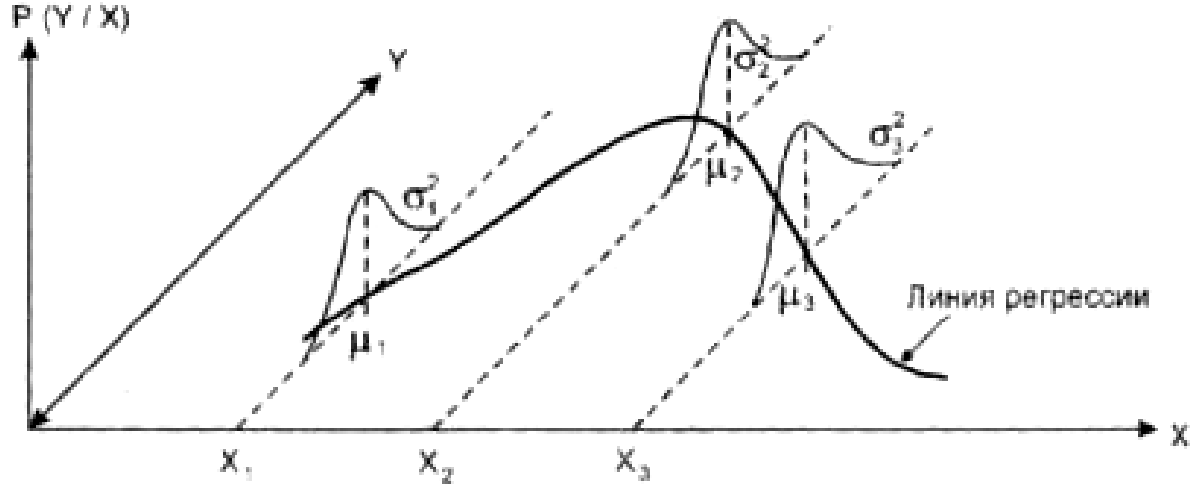


Рис. 25. Статистические предположения, лежащие в основе регрессионного анализа

Условные распределения зависимой переменной Y нормальны. Их математические ожидания μ_1, μ_2, μ_3 лежат на линии регрессии; дисперсии $\sigma_1^2, \sigma_2^2, \sigma_3^2$ равны. При этом μ_1, μ_2, μ_3 – математические ожидания тех условных распределений переменной Y, которые получаются при фиксации значений соответственно X_1, X_2, X_3 переменной X. Линия регрессии говорит о том, насколько статистически изменится среднее значение Y при изменении значения X. Точность, с которой линия регрессии Y по X передает изменение Y в среднем при изменении X, измеряется дисперсией величины Y для каждого X:

$$D\,(Y/X) = s^2\,(X).$$

Пусть $\sigma_1^2, \sigma_2^2, \sigma_3^2$ значения дисперсий, вычисленных для условных распределений переменной Y, получающихся при фиксации значений соответственно 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

рис. 26. При равенстве дисперсий говорят, что условные распределения удовлетворяют свойству **гомоскедастичности.** Чем меньше условные дисперсии Y, т. е. чем меньше разброс зависимого признака в условных распределениях, тем более достоверен регрессионный прогноз. Большой разброс снижает его достоверность.

Линия регрессии обладает свойством: **среди всех действительных функций f минимум математического ожидания $\mu(Y-f(X))^2$ достигается для функции f (X) = μ(Y/X).** Поясним это положение по рис. 26.

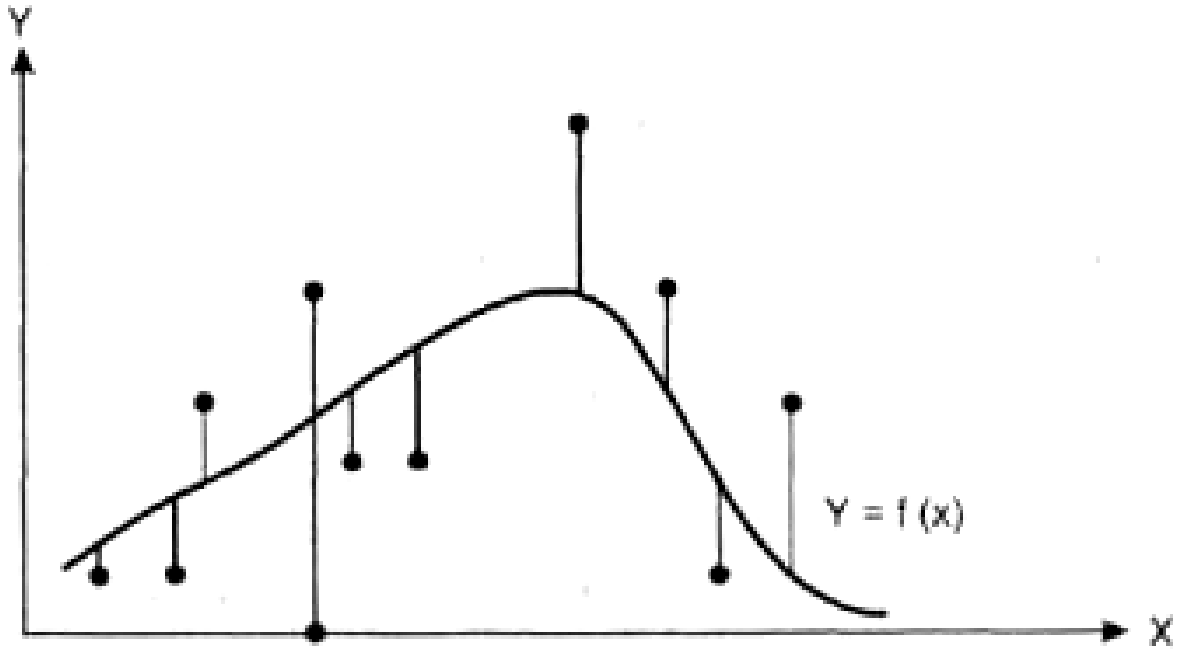


Рис. 26. Отклонения ординат рассматриваемых точек от произвольной функции

Вертикальные отрезки – отклонения ординат рассматриваемых точек от графика 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 рис. 27.

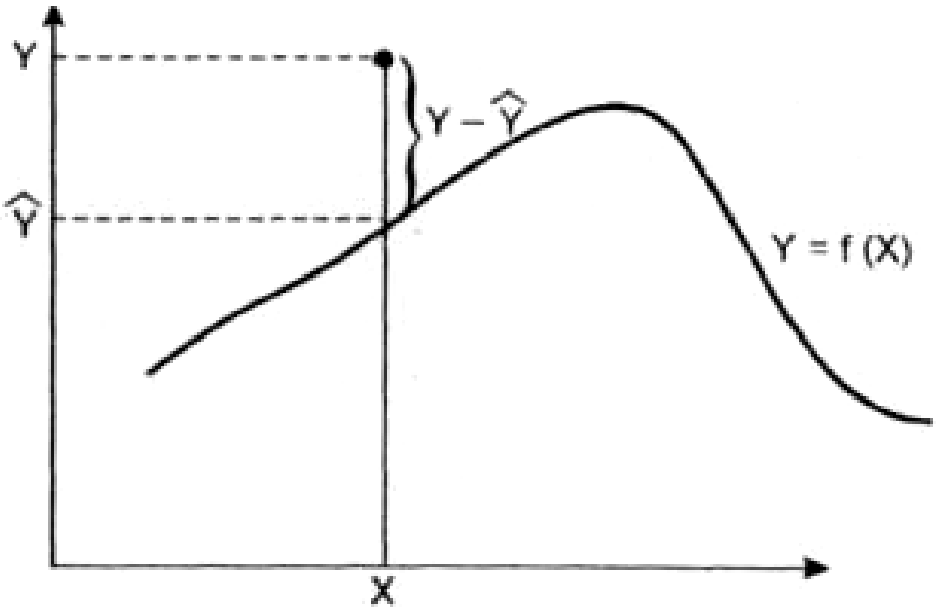


Рис.27. Отклонение точки (X, Y)
от произвольной функции Y = f (X)

X координата рассматриваемого объекта по оси X; Y – координата по оси Y; $\hat{Y}$ – ордината точки, принадлежащей графику функции Y = f (X) и имеющей по оси X ту же координату, что и объект.

Сумма $\sum (Y - \hat{Y})^2$ и есть та величина, которую надо минимизировать для того, чтобы получить выборочное представление линии регрессии. При этом суммирование осуществляется по всем рассматриваемым объектам:

$$\sum (Y - \hat{Y})^2 \rightarrow \min,$$

где $\hat{Y}$ – теоретическое, модельное значение зависимой переменной.

Минимальной эта сумма будет, если рассматриваемая функция Y = f (X) является выборочным представлением искомой линии регрессии. Чтобы найти выборочную линию регрессии, необходимо перебрать все возможные функции Y= f (X), для каждой вычислить указанную сумму квадратов и остановиться на той функции, для которой эта сумма минимальна. Этот способ поиска f (X) называется **метод наименьших квадратов,** и он задействован в широко применяемом в социологии методе парных сравнений.

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

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

где $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ – некоторые параметры, выборочные оценки которых надо получить. Однако найденная функция будет очень сложной и прогнозировать с ее помощью трудно. Поэтому выбирают какое-либо семейство кривых, имеющих сравнительно простые формулы, 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, все они 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.***

Если мы наблюдаем многомерный случай, т.е. такую ситуацию, когда имеется много независимых переменных $X_1, X_2, ..., X_n$ ($n > 1$), то сказанное выше также справедливо. Отличие только в том, что линейная регрессионная модель имеет вид не прямой линии, а гиперплоскости:

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_nX_n.$$

Уравнение регрессии будет более ясным, если все эти единицы будут одинаковыми. Для этого обычно осуществляют так называемую стандартизацию всех значений каждого признака (**нормировку**): вычитают из каждого такого значения среднее арифметическое признака и делят полученную разность на его дисперсию.

Рассмотрим признак X_2. Если X_2^1 – некоторое (i-е) его значение, $\bar{X}_2$ и s_X – отвечающие ему среднее арифметическое и дисперсия, то нормировка будет означать преобразование значения X_2^1:

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

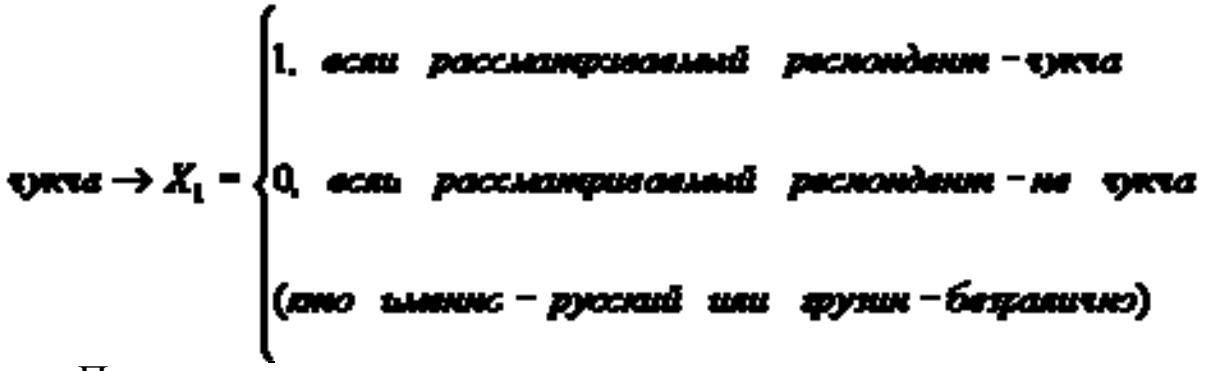
Для того чтобы на основе информации, полученной по номинальной шкале, можно было построить уравнение регрессии, эту информацию необходимо преобразовать – **дихотомизировать номинальные данные.** Вместо каждого номинального признака, принимающего k значений, вводим k' новых дихотомических (0 и 1).

Предположим, что рассматриваемый номинальный признак X – это национальность и что в закрытом вопросе анкеты названы три национальности: русский, грузин и чукча. Дихотомизируем их[73]:

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

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

[73] Толстова Ю.Н. Указ. соч. С. 168.



Предположим, что мы хотим изучить связь вида

$$Y = f(X),$$

где X – национальность, а Y – профессия. Вместо признака X в уравнение необходимо вставить X_1, X_2, X_3. Однако нежелательно включать в регрессионную модель такие предикторы, которые заведомо связаны друг с другом. А относительно наших X_1, X_2, X_3 такая связь есть. Как поступить в данном случае?

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 определим ее методом исключений, если знаем, какие значения для него имеют признаки X_1 и X_2 (табл. 34).

Таблица 34

Зависимость 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

Один дихотомический признак как бы отбрасывают, и число аргументов уравнения будет на единицу меньше, чем число альтернатив в номинальном признаке. В нашем случае вместо трех мы включаем в уравнение только два (отбросили X_3).

Теперь рассмотрим ситуацию с зависимой переменной Y, которая по анкете также имеет несколько дихотомических признаков: учитель, торговец, дворник[74].

$$Y_1 \begin{cases} 1, & \textit{если респондент – учитель} \\ 0, & \textit{если респондент – не учитель} \end{cases}$$

$$Y_1 \begin{cases} 1, & \textit{если респондент – торговец} \\ 0, & \textit{если респондент – не торговец} \end{cases}$$

$$Y_1 \begin{cases} 1, & \textit{если респондент – дворник} \\ 0, & \textit{если респондент – не дворник} \end{cases}$$

Строим три уравнения регрессии, каждое из которых отвечает своему Y_i:

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

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

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

Допустим, имеются некоторые номинальные признаки Y и X_1, X_2, ..., X_n. 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_1^1, X_2^1, ..., $X_{l_1}^1$. Отбрасываем последний признак из набора. Применение регрессионного анализа означает расчет k уравнений вида:

$$Y_1 = f_1(X_1, X_2,..., X_n) =$$
$$= f_1(X_1^1, X_2^1,..., X_{l_1-1}^1, X_1^2, X_2^2, ..., X_{l_2-1}^2, ..., X_1^n, X_2^n, ..., X_{l_n-1}^n);$$

[74] Толстова Ю.Н. Указ. соч. С. 168.

$$Y_2 = f_2(X_1, X_2, ..., X_n) =$$
$$= f_2(X_1^1, X_2^1, ..., X_{k-1}^1, X_1^2, X_2^2, ..., X_{k-1}^2, ..., X_1^n, X_2^n, ..., X_{k-1}^n)\ldots$$
$$Y_k = f_k(X_1, X_2, ..., X_n) =$$
$$= f_k(X_1^1, X_2^1, ..., X_{k-1}^1, X_1^2, X_2^2, ..., X_{k-1}^2, ..., X_1^n, X_2^n, ..., X_{k-1}^n).$$

Искомая зависимость имеет вид:

$$Y = f(X_1, X_2) = a_0 + a_1 X_1 + a_2 X_2.$$

Коэффициенты уравнения регрессии, найденные по правилам классического регрессионного анализа, выражаются сложными формулами, включающими в себя такие (неприемлемые для номинальных данных) статистики, как среднее арифметическое, дисперсия, частные коэффициенты корреляции и т.д. Однако социолог может рассмотреть их как условные частоты. Интерпретируем a_0, a_1, a_2.

Коэффициент a_0. Рассмотрим только тех людей, которым соответствует отброшенная нами национальность – чукчей.

$$X_1 = X_2 = 0.$$

Подставив эти значения в уравнение регрессии, получим соотношение

$$Y = a_0,$$

где a_0 равен среднему арифметическому значению зависимой переменной для отброшенной категории респондентов и означает долю чукчей, работающих торговцами.

Коэффициент a_1. Рассмотрим только русских. $X_1 = 1$ и $X_2 = 0$. Подставим эти значения в уравнение:

$$Y = a_0 + a_1,$$

где 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:

$$Y= 0,3 – 0,1\ X_1 + 0,6\ X_2.$$

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.

Приведем еще один пример[75]: пусть X – семейное положение (X_1 – женат, X_2 – неженат), Y – посещение кинотеатра (Y_1 – посещает, Y_2 – не посещает). Пусть таблица сопряженности, отвечающая данным признакам, имеет вид табл. 35:

Таблица 35

Схематическое изображение таблицы сопряженности для признаков X – семейное положение, Y – посещение кинотеатра

Значения Y	Значения 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 $Y = \alpha + \beta X$:

$$\alpha = \frac{b}{b+d},$$

где α – доля посещающих кинотеатр среди неженатых;

$$\beta = \frac{a}{a+c} - \frac{b}{b+d},$$

Пусть матрица имеет вид четырехклеточной таблицы сопряженности (табл. 36):

[75] Типология и классификация в социологических исследованиях. С. 260 – 266.

Таблица 36

Матрица сопряженности для признаков X – семейное положение, Y – посещение кинотеатра

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

Тогда:

$$\alpha = \frac{b}{b+d} = \frac{38}{50} = 0,76;\ \beta = \frac{a}{a+c} - \frac{b}{b+d} = 0,96 - 0,76 = 0,2.$$

Следовательно, Y = 0,76 + 0,2X.

Коэффициенты уравнения регрессии – более важная альтернатива обычному частному распределению, используемому социологом.

Таким образом, **с помощью НРА мы можем решать несколько типов задач:**

1. Нахождение определенных условных процентов с одновременным получением возможности прогноза.

2. Осуществление поиска взаимодействий (см. алгоритмы типа THAID и CHAID).

3. Осуществление сложных прогнозов.

6.5.2. Логит- и пробит-модели

Класс решаемых с помощью техники номинального регрессионного анализа задач может быть расширен за счет использования логистической регрессии, логит-моделей.

Линейное регрессионное уравнение чаще всего имеет следующий вид:

$$m = a + b_1X_1 + b_2X_2 + : + b_kX_k.$$

Принято называть **связующей функцией** такую функцию g, для которой справедливо соотношение

$$g(m) = a + b_1x_1 + b_2x_2 + : + b_kx_k.$$

Если g – тождественная функция (g(m) = m, identity link), то соотношение превращается в обычную регрессию. Если же g – это логарифм (log link), то мы получаем **логлинейную модель:**

$$\log(m) = a + b_1x_1 + b_2x_2 + : + b_kx_k.$$

Преимущество использования логлинейной модели в том, что она дает возможность свести изучение сложных взаимодействий между независимыми переменными (т.е. подбор таких произведений x, которые делают модель адекватной реальности) к поиску коэффициентов линейной зависимости (поскольку логарифм произведения равен сумме логарифмов). Особую важность для социолога имеет т.н. **логит-связь**, когда функция g является функцией вида:

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

Эта модель играет большую роль, когда Y – дихотомическая переменная. Если p – доля единичных значений Y, а доля нулевых значений q = (1-p), то

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

где функция g является логарифмом отношения преобладания.

Пусть у нас только один признак X. Тогда уравнение вида

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

называется **логистической регрессионной функцией.**

Не менее важна и т. н. **линейная вероятностная модель**

$$P(X) = a + bx.$$

Если 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.

Глоссарий

Анализ детерминационный – система методов анализа социологических данных, в которых задачи обработки и интерпретации рассматриваются как условное объяснение одного свойства посредством другого.

Анализ дискриминантный – вид статистического многомерного анализа, в котором при наличии нескольких генеральных совокупностей и их выборок требуется построить максимально эффективное классифицирующее правило, позволяющее приписать новый элемент генеральной совокупности.

Анализ дисперсионный – метод математической статистики, предназначенный для выявления влияния отдельных независимых друг от друга признаков-факторов на некий наблюдаемый признак.

Анализ ковариационый – совокупность методов математической статистики, относящихся к анализу моделей зависимости среднего значения некоей случайной величины от набора неколичественных факторов и одновременно от набора количественных факторов.

Анализ латентно-структурный – метод вероятностно-статистического моделирования, предполагающий, что ответы респондентов на вопросы есть внешнее проявление скрытой латентной характеристики. Суть метода в открытии характеристики и классификации ее носителей.

Анализ лонглинейный – статистический метод изучения многомерных таблиц сопряженности. Позволяет статистически проверить гипотезу о системе одновременно имеющих место

парных и множественных взаимосвязей в группе признаков, измеренных по номинальным шкалам.

Анализ причинный (путевой) – методы моделирования причинных отношений между признаками с помощью систем статистических уравнений, чаще регрессионных.

Анализ регрессионный – статистический метод исследования зависимости (регрессии) между зависимым признаком Y и независимыми X (регрессорами, предикторами).

Анализ регрессионный качественный – группа методов многомерного анализа данных, позволяющих оценить влияние нескольких качественных (классификационных или номинальных) независимых признаков-предикторов на зависимый признак Y (регрессионный анализ с дихотомическими переменными, множественный классификационный анализ, множественный номинальный анализ и пр.).

Анализ типологический – метод изучения сложных социальных объектов, состоящий в выделении социально значимых отличных и внутренне однородных групп объектов , характеризующихся совокупностью признаков произвольной природы. Осуществляется путем операционализации понятий и формализации с помощью математических методов.

Этапы А.Т.:

1. Построение априорной типологии.
2. Определение объекта типологии.
3. Операционализация.
4. Построение признакового пространства.
5. Выбор формального аппарата классификации.
6. Определение стратегии интерпретации.

Анализ факторный – группа методов многомерного статистического анализа, которые компактно представляют обобщенную информацию о структуре связей между наблюдаемыми при-

знаками социального явления на основе некоторых скрытых факторов, непосредственно не наблюдаемых.

Анализ факторный количественных данных – группа методов многомерного анализа, для данных, полученных по измерениям по интервальным шкалам, где каждый наблюдаемый признак можно выразить в виде суммы некоторых других, ненаблюдаемых признаков, умноженных каждый на свой коэффициент (факторную нагрузку).

Анализ факторный качественных данных – группа методов многомерного анализа, базирующихся на предположениях традиционного факторного анализа, применяемого к качественным категоризированным данным.

Корреляция – зависимость между числовыми случайными величинами, которая в отличие от функциональной зависимости рассматривается тогда, когда по крайней мере одна из величине зависит не только от другой, но и от ряда случайных факторов.

Коэффициенты парной связи номинальных признаков – коэффициенты, которые показывают наличие статистической связи между признаками:

– коэффициенты – показатели существования связи (хи-квадрат, показатель средней квадратичной сопряженности, коэффициент Пирсона, коэффициент Чупрова, коэффициент Крамера);

– коэффициенты – показатели прогноза значений одного признака по значениям другого:

а) коэффициенты, свидетельствующие об уменьшении ошибки предсказания (коэффициент Гудмена-Краскала, коэффициент Гуттмана);

б) коэффициенты информационных мер связи (ассиметричный коэффициент неопределенности, показатель влияния, коэффициент Райского);

– коэффициент близости разбиений.

Коэффициенты ранговой корреляции – выборочные меры зависимости 2-х случайных величин, основанные на ранжировании независимых результатов наблюдений; наиболее известны коэффициенты Спирмена и Кендалла.

Показатели корреляции – коэффициенты, выражающие силу корреляции между числовыми случайными величинами. Наиболее известны:

– ковариация – числовая характеристика совместного распределения 2-х случайных величин, равная математическому ожиданию произведения отклонений случайных величин от их математических ожиданий;

– коэффициент корреляции – числовая характеристика совместного распределения 2-х случайных величин, выражающая их взаимосвязь;

– корреляционное отношение (для нелинейных зависимостей) – характеристика отношений зависимости между случайными величинами.

Меры рассеяния – статистические показатели, характеризующие степень разброса значений признака относительно среднего значения (для признаков количественного характера), или равномерного распределения (для признаков номинального типа). Наиболее известны:

– дисперсия;

– коэффициент вариации;

– среднее абсолютное отклонение.

Для признаков номинального характера возможно применение:

– коэффициента энтропии;

– коэффициента качественной вариации.

Шкала – алгоритм, с помощью которого осуществляется измерение в тех случаях, когда объекты переводятся в числовую математическую систему путем присвоения числа, называемого шкальным значением объекта.

Шкала Гуттмана – основа шкалограммного анализа для обработки данных, образованных ответами на вопросы типа «да – нет». Метод предполагает сочетание шкалы вопросов и шкалы респондентов; кроме того, это дополняется системой числовых индексов, позволяющих оценить, насколько данные согласуются с моделью.

Шкала Лайкерта – метод шкалирования социально-психологических характеристик индивидов. Согласие или несогласие с суждением оценивается по 5 – 7-балльной шкале. Выведение суммарной оценки суждений является вариантом анализа суждений.

Шкала Терстоуна – метод шкалирования социально-психологических характеристик индивидов, основанный на предварительном измерении шкальных значений набора суждений, отражающих различную степень выраженности измеряемой характеристики (от 15 до 30), которые полностью покрывают спектр изучаемой установки или ценности. Эталонирование суждений происходит с использованием методов равнокажущихся интервалов, последовательных интервалов и парных сравнений.

Вопросы к зачету

1. Статистические закономерности в анализе социологической информации.
2. Моделирование социальной реальности.
3. Специфика математико-статистических методов применительно к социологической информации.
4. Задачи математики применительно к социологической информации.
5. Сложности использования математики в социологии.
6. Процедура анализа данных в социологии.
7. Этапы анализа данных.
8. Виды анализа данных в социологии.
9. Специфика анализа одномерных распределений.
10. Шкалирование и виды шкал.
11. Методы анализа одномерных распределений.
12. Дисперсия и среднеквадратическое отклонение.
13. Меры средней тенденции.
14. Меры вариации.
15. Энтропийный коэффициент разброса.
16. Специфика анализа двумерных распределений.
17. Роль номинальных данных в социологии.
18. Таблицы сопряженности.
19. Анализ связей между номинальными признаками: общая характеристика.
20. Коэффициенты корреляции.
21. Коэффициенты критерия *хи-квадрат*.

22. Коэффициенты связи, основанные на различных моделях прогноза.
23. Коэффициенты связи, основанные на понятии энтропии.
24. Коэффициенты связи для четырехклеточных таблиц сопряженности.
25. Многомерные отношения преобладаний.
26. Детерминационный анализ.
27. Анализ фрагментов таблиц.
28. Алгоритмы поиска сочетаний независимых предикатов.
29. Нелинейный регрессионный анализ и его применение в социологии.
30. Логит- и пробит-модели в социологии.

Библиография

1. Agresti, A. Categorical data analysis / A.Agresti. – N.-Y.: John Wiley and sons, 1990.

2. Clausen, S.-E. Applied correspondence analysis. An introduction. Sage university paper series on Quantitative applications in the social sciences, 07-121 / S.-E. Clausen. – Newbury park, CA: Sage, 1998.

3. Rudas, T. Odds ratios in the analysis of contingency tables. Sage university paper series on Quantitative applications in the social sciences, 07-119 / T. Rudas. – Newbury park, CA: Sage, 1998.

4. Аптон, Г. Анализ таблиц сопряженности / Г.Аптон. – М.: Финансы и статистика, 1982 (Upton G.J.G. The analysis of cross-tabulated data. N.-Y.: J.Wiley&Sons, 1978).

5. Гнеденко, Б.В. Курс теории вероятностей / Б.В. Гнеденко. – М.: Наука, 1965.

6. Добреньков, В.И. Методы социологического исследования / В.И. Добреньков, А.И. Кравченко. – М.: Инфра-М, 2006.

7. Дэйвисон, М. Многомерное шкалирование / М. Дэйвисон. – М.: Финансы и статистика, 1988.

8. Елисеева, И.И. Статистические методы измерения связей / И.И.Елисеева. – Л.: ЛГУ, 1982.

9. Елисеева, И.И. Группировка, корреляция, распознавание образов / И.И. Елисеева, В.О. Рукавишников. – М.: Статистика, 1977.

10. ДА-система (Детерминационный анализ). – М.: Контекст. – 1997.

11. Кендалл, М.Дж. Статистические выводы и связи / М.Дж. Кендалл, А. Стьюарт. – М.: Наука, 1973.

12. Миркин, Б.Г. Анализ качественных признаков и структур / Б.Г. Миркин. – М.: Статистика, 1980.

13. Миркин, Б.Г. Группировки в социально-экономических исследованиях / Б.Г. Миркин. – М.: Финансы и статистика, 1985.

14. Паниотто, В.И. Количественные методы в социологических исследованиях / В.И. Паниотто, В.С. Максименко. – Киев: Наукова Думка, 1982.

15. Рабочая книга социолога. – М.: Наука, 1983.

16. Социология: Словарь-справочник. Т. 4. Социологическое исследование: методы, математика и статистика. – М., 1991.

17. Толстова, Ю.Н. Математика в социологии: элементарное введение в круг основных понятий (измерение, статистические закономерности, принципы анализа данных) / Ю.Н. Толстова. – М.: ИСАН СССР, 1990.

18. Толстова, Ю.Н. Измерение в социологии / Ю.Н. Толстова. – М.: Инфра-М, 1998.

19. Толстова, Ю.Н. Анализ социологических данных / Ю.Н. Толстова. – М.: Научный мир, 2000.

20. Философский энциклопедический словарь. – М.: Наука, 1983.

21. Чесноков, С.В. Детерминационный анализ социально-экономических данных / С.В. Чесноков. – М.: Наука, 1982.

22. Яглом, А.М. Вероятность и информация / А.М. Яглом, И.М. Яглом. – М.: Гос. Изд-во физ-мат. литературы, 1960.

23. Ядов, В.А. Стратегия социологического исследования: описание, объяснение, понимание социальной реальности / В.А. Ядов. – М.: Добросвет, 1998.

24. Яшин, В.П. Корреляционный анализ в социологических и психологических исследованиях / В.П. Яшин. – Н. Новгород: Изд-во НКИ, 1999.

Рекомендуемая зарубежная литература

1. Bluman, A.G. Elementary statistic / A.G. Bluman. – W.C. Brown Publishers. 1995.

2. Clausen, S.-E. Applied correspondence analysis. An introduction. Sage university paper series on Quantitative applications in the social sciences, 07-121 / S.-E. Clausen. – Newbury park, CA: Sage, 1998.

3. Demaris, A. Logit modeling: Practical application. Sage university paper series on quantitative applications in the social sciences, 07-086 / A. Demaris. – Newbury park, CA: Sage, 1992.

4. Magidson, J. The CHAID approach to segmentation modeling / J. Magidson // Handbook of marketing research. – Cambridge, Mass.: Blackwell, 1993.

5. McCutcheon, A.L. Latent class analysis. Sage university paper series on quantitative applications in the social sciences, 07-064 / A.L. McCutcheon. – Newbury park, CA: Sage, 1987.

6. Menard, S. Applied logistic regression analisys. Sage university paper series on Quantitative applications in the social sciences, 07-106 / S. Menard. – Newbury park, CA: Sage, 1995.

7. Morgan, J.N. THAID – a sequential analysis program for nominal dependent variables / J.N. Morgan, R.C. Messenger. – Ann. Arbor: Institute for social research, 1973.

Содержание

Учебное издание

Епархина Ольга Валерьевна

Математические методы обработки и анализа социологических данных

Учебное пособие

Редактор, корректор О.Н. Скибинская
Компьютерная верстка И.Н. Ивановой

Подписано в печать 26.02.2007. Формат 60x84/16. Бумага тип.
Усл. печ. л. 7,67. Уч.-изд. л. 5,19. Тираж 60 экз. Заказ

Оригинал-макет подготовлен
в редакционно-издательском отделе ЯрГУ
Ярославский государственный университет
150000 Ярославль, ул. Советская, 14

Отпечатано
ООО «Ремдер» ЛР ИД № 06151 от 26.10.2001.
г. Ярославль, пр. Октября, 94, оф. 37
тел. (4852) 73-35-03, 58-03-48, факс 58-03-49.

132