ГЛАВА 3

ВОСХОДЯЩАЯ СТРАТЕГИЯ АНАЛИЗА ДАННЫХ

1. С ЧЕГО НАЧИНАЕТСЯ АНАЛИЗ?

Восходящая стратегия анализа и нисходящая стратегия анализа. Различие понятий «анализ данных» и «логика анализа» (логическая схема анализа). Первичный анализ как составная часть любой стратегии. Признак. Анализ «поведения» отдельно взятого признака. Вариационный ряд. Одномерное распределение. Показатели распределения. Абсолютная, относительная и накопленная частоты. Деление на интервалы. Цели первичного анализа данных. «Язык» анализа распределений.

Следует напомнить, что в качестве третьего структурного элемента области эмпирической социологии, обозначенной нами как ***методология анализа данных,*** выделена:

восходящая стратегия анализа (проверки описательных гипотез; поиск эмпирических закономерностей, начиная с простых и заканчивая сложными для формирования новых гипотез).

Следует особо остановиться на использовании пары понятий: ***восходящая стратегия анализа - нисходящая стратегии анализа.*** Что касается просто пары понятий «восходящая стратегия - нисходящая стратегия», то она используется в эмпирической социологии в разных контекстах. Например, для обозначения выборочной стратегии в исследовании. Если сбор информации осуществляется по так называемому методу «снежного кома», то это пример восходящей выборочной стратегии. Такая стратегия используется обычно для изучения латентных социальных групп (наркоманов, скрытых алкоголиков и т. д.). С нисходящей выборочной стратегией мы сталкиваемся при формировании выборки, исходя из структуры генеральной совокупности. Это является типичным для изучения общественного мнения. Разумеется, в рамках одного и того же исследования одновременно могут использоваться как нисходящая, так и восходящая стратегии формирования выборки.

Такую пару терминов можно использовать и для характеристики логики исследовательского процесса в целом, а именно для обозначения двух подходов к изучению социальной реальности. Мы их обозначили как статистическую и гуманитарную традиции (подходы, парадигмы) в эмпирической социологии. Известно, что латентные социальные группы, в отличие от других, целесообразнее изучать по восходящей стратегии [6], т. е. не только стратегия формирования выборки носит восходящий характер, но и все исследование в целом построено по восходящей стратегии изучения таких групп.

Эта пара терминов применяется и в достаточно узком смысле в так называемых методах многомерной классификации для обозначения процедуры деления эмпирических объектов на группы. О понятии «классификация» пойдет речь в последней части книги. Это пока ремарка для «всезнаек». Пара «восходящая стратегия анализа

данных - нисходящая стратегия анализа данных» составляет ***основу*** для формирования в социологическом исследовании ***логики анализа данных, логической схемы анализа.*** Социолог выбирает стратегию анализа данных исходя из специфики своего исследования (цели, задачи, гипотезы). Рассмотрим несколько исследовательских ситуаций.

Первая ситуация

Предположим, у социолога нет четко обозначенных гипотез ни описательного, ни объяснительного характера. Разумеется, в расплывчатой форме они всегда существуют. Ведь социолог, желая «взглянуть» на социальную реальность через призму какого-то подхода, отвечает на вопросы ***«Что изучать?»*** **и** ***«Зачем и для достижения каких целей изучать?».*** Отсутствие четкости в гипотезах требует определенной стратегии при работе с эмпирическим материалом. Сначала социолог в «мешке» с информацией наводит «косметический» порядок - ищет простые эмпирические закономерности. Их можно назвать и ***регулярностями.*** Прежде всего он выделяет сами эмпирические индикаторы, если их нет. С этой ситуацией мы сталкиваемся при работе с текстовой информацией. Мы с вами выделяли элементарные обоснования и элементы в контексте применения метода неоконченных предложений. Первые из них и являлись эмпирическими индикаторами.

При работе с биографиями людей, с текстами полуформализованных и свободных интервью естественным образом появляется необходимость в анализе, условно говоря, *«поведения»* отдельно взятого эмпирического индикатора. Затем возникает потребность в анализе совместного «поведения» двух эмпирических индикаторов, в анализе их взаимосвязей. Таким образом, логика анализа эмпирии строится по восходящей (от частного к общему) стратегии. Начальный этап такой стратегии - ***первичный анализ / первичная обработка данных.***

Социолог, исходя из восходящей стратегии, последовательно ищет ответы на вопросы, такие, как: не объединяются ли эмпирические индикаторы в какие-то группы, а объекты - в классы. К примеру, похожие в определенном смысле объекты представляют собой некий класс, а взаимосвязанные между собой эмпирические индикаторы могут образовать некую группу. Вполне возможно, что объекты, отнесенные к одному и тому же формальному классу, являются однотипными. А группа эмпирических индикаторов может интерпретироваться как некий специфический социальный фактор. О содержании понятий «тип» и «фактор» пойдет речь в последней главе. Главная задача в таких исследовательских сюжетах - проблема ***интерпретации*** разного рода эмпирических закономерностей, ибо они выражают какие-то тенденции, синдромы.

Вторая ситуация

У социолога могут быть четко обозначены гипотезы исследования. В этом случае логика анализа может строиться как в рамках восходящей, так и нисходящей стратегий. Выбор стратегии зависит от характера гипотез и от того, какими априорными знаниями (знания, имевшиеся до проведения исследования) располагает исследователь. Допустим, что источником эмпирической информации является индивид; техника сбора данных жестко структурирована; в исследовании проверяются только описательные гипотезы. Тогда также необходимы восходящие, от частного к общему, стратегии анализа. Вспомним из предыдущего материала, что в процессе прямого ранжирования для принятия решения о присвоении рангов нам непременно требовалось изучить степень единодушия респондентов в оценке объектов ранжирования. Для этих целей в процессе анализа опять же требуется движение по восходящей стратегии.

Пусть гипотеза звучит следующим образом: политические пристрастия населе-

ния в основном определяются возрастом и происхождением. Для проверки этой гипотезы социолог определяет всевозможные связи этих »пристрастий» с огромной совокупностью различных эмпирических индикаторов. Если из всех связей оказываются самыми сильными связь с возрастом и с происхождением, то считается, что гипотеза подтвердилась. К примеру, сформулируем другую гипотезу: в России существуют типы электорального поведения областей, интерпретируемые как объекты социального управления. В том смысле, что механизм воздействия на отдельные области одинаков, если они отнесены к одному и тому же типу. Для проверки такой сложной гипотезы необходимую основу для логики анализа составляет нисходящая стратегия анализа (от общего к частному). Такой пример будет приведен в последней главе.

Ясно одно: проверка такого рода гипотез предполагает «продумывание» всей логики анализа априори (до сбора эмпирической информации). Это очень не просто. Вместе с тем такое «продумывание» нужно и важно даже в описательных исследованиях. А в серьезных аналитических исследованиях для проверки сложных гипотез тем более. Вспомним, что мы с вами рассматривали несложные модели изучения отдельных свойств социальных объектов для перехода с теоретического уровня на эмпирический. При этом совершенно не затрагивали вопросы обратного перехода, для которого крайне важно понятие логической схемы анализа.

Если вернуться к модели изучения свойства социального объекта, то, в контексте наших рассуждений, логика анализа позволяет уточнить не только саму такую модель, но и предполагает продумывание заранее логики получения эмпирических закономерностей и, соответственно, переход от них к теоретическим обобщениям. Разумеется, речь идет уже о сложных эмпирических закономерностях, получаемых на основе всей системы изучаемых в исследовании свойств. В зависимости от логической схемы анализа социолог определяет и то, какого рода эмпирический материал ему нужен, и то, какие приемы «обработки» информации необходимы, и то, в какой последовательности будет строиться логика изучения и ***объяснения*** того или иного социального феномена. В таких исследовательских сюжетах главным является концептуальная схема, теория «видения» социальной реальности, так как идет поиск ответа на вопрос ***«Почему это?».*** Для такого случая необходима нисходящая (от общего к частному) стратегия анализа. Поиск ответа на вопрос «Почему это?», проверка объяснительных гипотез социологического исследования возможны только в рамках нисходящей стратегии анализа. Все, что с этим связано, будет обсуждаться в последней части книги.

В отдельно взятом социологическом исследовании возможно сочетание восходящей и нисходящей стратегий анализа. Та и другая стратегии могут быть реализованы на практике с помощью одних и тех же методов, приемов, способов «обработки» информации. Например, к таковым относятся так называемые методы математической статистики (это такая область математической науки, которая в определенной мере как бы обслуживает науки, работающие с эмпирическим материалом) и методы многомерного анализа. Сюда включаются и такие методы, применение которых теоретически может быть необоснованно. В том смысле, что закономерности, полученные для выборки, нельзя распространить (перенести) на всю генеральную совокупность. Однако эти методы «хорошо» работают на практике и их принято называть эвристическими в отличие от статистических. К различию понятий «статистика» и «эвристика» мы еще вернемся. Вся совокупность технических приемов (по сути, это использование математического формализма или математических методов в социологии) называется методами ***анализа данных.***

К этому разделу мы подошли с пониманием того, что социологу, изучающему различные социальные феномены, приходится строить модели изучения их свойств, пользоваться различными типами информации, применять совокупность приемов измерения латентных, непосредственно не наблюдаемых признаков, выбирать стра-

тегию анализа. Это и есть начало начал анализа данных.

Наблюдаемые признаки мы называли эмпирическими индикаторами. В предыдущих разделах они были нашими главными понятиями. Здесь и далее таковыми будут ***признаки.*** Признаком может быть и отдельно взятый эмпирический индикатор, и производный от них показатель. Например, признаком будем называть любые показатели, индексы, коэффициенты, возникающие в рамках работы с данными типа «государственная статистика», «бюджет времени». Признак, как и любой эмпирический индикатор, имеет для нас те же три уровня измерения: номинальный, порядковый, «метрический». Как минимум, мы должны научиться изучать «поведение» всех трех типов признаков, измеренных по трем типам шкал.

Представляется важным еще раз повторить следующее. Несмотря на многообразие шкал (в данном случае как линеек для измерения чего-то) в социологии, мы рассматриваем только три типа шкал и к «метрическим» относим все шкалы, уровень измерения по которым выше порядкового, т.е. то, что очень похоже на числа, на «количества».

С чего же начинается анализ ***«поведения»*** отдельно взятого признака тогда, когда информация «лежит» на столе социолога? Такой анализ необходим практически всегда независимо от исследовательских задач, типов информации, выбора стратегии анализа. Речь идет как бы о «социальной бухгалтерии», азы которой вы должны освоить. Практически в любой книге, в название которой входят слова *«...статистические методы в...»,* вы найдете определенный материал по освоению этих азов [2, 3, 7, 8, 9, 11].

Несмотря на то, что ниже рассматривается пример, имеющий отношение к данным анкетирования, все выводы относятся к анализу любых вариационных и динамических рядов. К сожалению, объем *книги* не позволяет привести другие примеры. На протяжении всей этой главы в основном будем приводить фрагменты из некоторого исследования на тему «Структура времяпрепровождения студентов: сравнительный анализ вузов», придуманного (модельного) нами в качестве примера. Сбор данных осуществлялся в нем как по использованию бюджета времени, так и по вопроснику «сложной структуры»; генеральная совокупность - студенты вузов России, Нас в этом исследовании будут интересовать только студенты-гуманитарии, т. е. некоторая подвыборка.

Рассмотрим всего три признака из этого исследования: будущую профессию студента-гуманитария, его удовлетворенность учебой и продолжительность времени на учебу. Относительно третьего признака нужно подчеркнуть следующее. Продолжительность в данном случае представляет собой сумму затрат времени на прослушивание лекций, на участие в семинарских занятиях, на дополнительные самостоятельные занятия, а также на перерывы между аудиторными занятиями. В качестве примера будем рассматривать среднесуточную, например, за неделю, продолжительность. ***«Продолжительность»*** имеет метрический уровень измерения. ***«Будущая профессия»*** как признак имеет номинальный уровень измерения. ***«Удовлетворенность учебой»*** может быть измерена посредством логического квадрата по пятибалльной порядковой шкале. Тогда она понимается только как удовлетворенность учебой в «родном» вузе (вернитесь к тому разделу, где обсуждается логический квадрат). Вместо этих признаков можно было бы выбрать и любые другие.

Что означает анализ «поведения» профессии на совокупности объектов? Это означает, что мы должны ***обработать*** эмпирические данные, чтобы получить ***распределение*** изучаемых объектов (в нашем случае студентов-гуманитариев) по профессиональным группам и по ***характеру*** этого распределения судить о профессиональной ***структуре*** опрошенных студентов. Для простоты изложения буду приводить цифры модельного характера, т. е. в реальном исследовании они не были получены. Предположим, что нас интересует восемь профессий, все они закодированы

цифрами от 1 до 8, а число студентов-гуманитариев среди всех опрошенных равно 1000. Таким образом, исходно мы имеем матрицу данных типа «объект - признак». Из нее выделяем для анализа столбец матрицы в соответствии с анализируемым признаком. Подсчитываем в этом ряду число респондентов, которые в недалеком будущем будут иметь ту или иную профессию. Тем самым получаем ***частоту*** встречаемости в выборке студента той или иной будущей профессии.

Распределение опрошенных по профессиям представлено в таблице 3.1.1. Это результаты самого первого этапа систематизации эмпирических данных. Распределение может быть представлено и описано на «языке» четырех показателей. Первый - ***абсолютная частота,*** т. е. число студентов с определенной «будущей» профессией. Среди опрошенных студентов оказалось 100 будущих политологов (профессия 1), 200 социологов (профессия 2), 300 культурологов (профессия 3), 100 филологов (профессия 4), 50 психологов (профессия 7) и 250 историков (профессия 8). Студенты с будущими профессиями, обозначенными как 5 и 6, в выборку не попали. В этом нет ничего удивительного, если при формировании выборочной совокупности не учитывалась будущая профессия студента. Эти шесть обозначенных и встречающихся в выборке профессий, будем использовать в процессе дальнейшего анализа.

Таблица 3.1.1

Распределение студентов по их будущей профессии

ПОКАЗАТЕЛИ	**БУДУЩАЯ ПРОФЕССИЯ СТУДЕНТА**								**Итого**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	
1. Абсолютная частота	100	200	300	100	-	-	50	250	1000
2. Относительная частота в долях (частость)	0,1	0,2	0,3	0,1	-	-	0,05	0,25	1
3. Относительная частота в %	10	20	30	10	-	-	5	25	100
4. Накопленная частота в %	НЕ ИМЕЕТ СМЫСЛА								

Второй показатель в таблице - ***относительная частота в долях,*** или ***частость,*** т. е. это доля респондентов определенной профессии среди всех ***опрошенных*** студентов-гуманитариев. Очень часто в социологических исследованиях наряду или вместо числа опрошенных используется число ***ответивших.*** Для нашего примера не имеет значения, по отношению к какому «числу» считается доля, ибо число ответивших совпадает с числом опрошенных. В массовых опросах различение этих величин носит принципиальный характер, так как число не ответивших бывает достаточно большим. Сама же ***проблема не ответивших*** является серьезной проблемой в массовых опросах. Мы касались этой проблемы при обсуждении так называемой (нами) проблемы социологического нуля. Относительная частота в долях - это важный показатель для последующих этапов работы с данными.

Доля интерпретируется как оценка вероятности обладать определенной профессией. Последняя фраза только для тех, кто случайно прослушал курс по теории вероятности.

Третий показатель - ***относительная частота в процентах*** — определяет, какой процент респондентов будет иметь ту или иную профессию. Это самый любимый показатель социолога, и вы в этом могли убедиться, если уже успели принять участие в каком-нибудь социологическом исследовании. ***Процент*** **и** ***частость*** - составные

элементы языка анализа социолога.

И, наконец, четвертый показатель - ***накопленная частота*** в процентах, С такой частотой мы сталкивались при построении шкалы Терстоуна. Для номинального уровня измерения она почти никогда не имеет смысла. Чисто технически ее можно подсчитать для нашей таблицы. Это и будет маленьким примером неадекватности математики. Прямо говоря - чушь. Отсюда и вывод, что, живя в век потрясающих компьютеров, слепо нажимать на кнопки для запуска «модерновых» математических методов недопустимо. Компьютер может подсчитать все, только есть ли в этом смысл. Вот в чем вопрос.

Накопленная частота имеет «прозрачный» содержательный смысл только для шкал начиная с порядковых. Рассмотрим распределение студентов по степени их удовлетворенности учебой, полученной с помощью применения логического квадрата. В таблице 3.1.2 представлено распределение респондентов по степени «удовлетворенности» по тем же четырем показателям (и в этом случае цифры не реальные, а модельные). Все показатели имеют смысл. Число опрошенных так же, как и в случае первого признака, совпадает с числом ответивших. Степени удовлетворенности обозначены цифрами от 1 до 5. При этом 1 соответствует минимальному уровню удовлетворенности, а 5 - максимальному.

Таблица 3.1.2

Распределение студентов по степени удовлетворенностью учебой

ПОКАЗАТЕЛИ	**СТЕПЕНЬ УДОВЛЕТВОРЕННОСТИ УЧЕБОЙ**					**Итого**
	1	**2**	**3**	**4**	**5**	
1. Абсолютная частота	200	300	200	250	50	1000
2. Относительная частота в долях (частость)	0,2	0,3	0,2	0,25	0,05	1
3. Относительные частоты в %	20	30	20	25	5	100
4. Накопленная частота	20	50	70	95	100	

Напомним, какой смысл имеет накопленная частота. Например, в таблице 3.1.2 частота, равная 70%, означает, что число студентов с уровнем удовлетворенности меньше четырех составляет 70% от числа опрошенных, а меньше трех - 50%. Перейдем к случаю метрической шкалы. Для табличного представления распределения «продолжительности» необходимо разбить диапазон ее изменения на отдельные интервалы. Важно отметить, что распределение не всегда имеет смысл представлять в табличной форме, так как деление на интервалы не всегда имеет смысл, например, для динамических рядов или для продолжительности затрат времени в исследованиях бюджета времени. Это происходит потому, что можно сразу переходить к изучению характеристик, описывающих характер распределения. Необходимо иметь также в виду, что признак может носить дискретный характер (встречаются только целые числа) или непрерывный характер (встречаются числа, имеющие целую часть и дробную). С непрерывностью встречаемся в основном при работе с аналитическими индексами на этапе анализа эмпирий.

Наш третий признак - продолжительность затрат времени на учебу - может носить дискретный характер, если выражен в минутах, и непрерывный характер - если выражен в часах. Остановимся на последнем случае. Для каждого студента этот про-

изводный показатель равен его среднесуточным (в часах) затратам времени на учебу. Введем интервалы и подсчитаем число студентов, внесенных в каждый интервал. В социологии в отличие от многих других наук, работающих с эмпирией, разбиение на интервалы не может носить формального характера. Такое разбиение всегда происходит в зависимости от исследовательских задач, а точнее, от того, как и для чего будет использоваться признак в процессе дальнейшего анализа. Поэтому социолог пользуется при этом понятиями «группировка данных», «типологическая группировка данных».

При выделении интервалов изменения продолжительности затрат времени на учебу исходим из значений максимальной и минимальной продолжительности, встретившихся в нашей выборке. Разница между этими величинами называется ***вариационным размахом.*** Без знания минимальной продолжительности нельзя определить нижнюю границу первого интервала, а без знания максимальной - верхнюю границу последнего интервала. Допустим, в нашем случае максимум (max) равен 9-ти часам, а минимум (min) - 0 часам. Последний факт можно объяснить тем, что в выборку попали студенты, которые были больны: никаких занятий, входящих в «продолжительность учебы», в недельном бюджете времени у них не было. Чтобы сей факт не вызвал недоумения, заметим, что сбор информации о бюджете времени студента происходит за неделю, предшествующую опросу.

Тогда наши интервалы (всего их шесть) могут выглядеть следующим образом:

1. 0—1 часов;
2. 1—2,5 часов;
3. 2,5—4 часов;
4. 4—7 часов;
5. 7—8 часов;
6. 8—9 часов.

Нетрудно догадаться, из чего мы исходили при выборе именно таких интервалов. К примеру, в последний интервал попадут студенты - «трудяги», в первый - те, кто по какой-то причине был «выключен» из учебного процесса, а в четвертый - модальная (самая распространенная) группа студентов. Кстати, это не факт, а гипотеза, и, соответственно, она может не подтвердиться в реальном исследовании. Для наглядности на рис 3.1.1 изображены эти интервалы в виде делений на линейке.

Рис 3.1.1

При отнесении респондента к конкретному интервалу по продолжительности учебы возникает такой вопрос. Куда входят нижняя и верхняя границы интервала? Другими словами, к какому интервалу отнести, например, студента, у которого продолжительность учебы равна четырем часам. Ведь его можно отнести и к первому, и ко второму интервалу. Эта проблема решается просто. Например, социолог принимает решение, что все верхние границы интервалов относятся к интервалу. Тогда студент, у которого продолжительность учебы равна 4-м часам, будет отнесен к третьему интервалу. Студент, у которого продолжительность учебы равна 8-ми часам, - к пятому и т. д.

Эти же интервалы могут быть заданы и в другой форме:

1. 0—1 часов;
2. 1,1—2,5 часов;
3. 2,6—4 часов;
4. 4,1—7 часов;

5. 7,1—8 часов;

6. 8,1—9 часов.

В этом случае при вычислениях возникает другая проблема, если продолжительность учебы некоторого студента, например, равна 1,09 часов. Опять же принятие решения в руках социолога. Он может отнести к интервалу не только верхнюю границу, но и то, что ниже нижней границы следующего интервала, т. е. респондент, у которого продолжительность учебы равна 1,09 часам будет отнесен к первому интервалу.

Используя первые введенные интервалы, подсчитаем по ним распределение респондентов (таблица 3.1.3.)

Таблица 3.1.3

Распределение студентов по продолжительности учебы

ПОКАЗАТЕЛИ	**ПРОДОЛЖИТЕЛЬНОСТЬ УЧЕБЫ**						**Итого**
	0-1	**1-2,5**	**2,5-4**	**4-7**	**7-8**	**8-9**	
1. Абсолютная частота	27	75	150	348	250	150	1000
2. Относительная частота в долях (частость)	0,027	0,075	0,15	0,348	0,25	0,15	1
3. Относительная частота в %	2,7	7,5	15	34,8	25	15	100
4. Накопленная частота в %	2,7	10,2	25,2	60	85	100	

Обратите внимание, что каждая из приведенных таблиц имеет заголовок, итоговый столбец. Эти таблицы - пример оформления как бы первичных результатов социологического исследования. Разумеется, за исключением того, что реальные таблицы содержат только один показатель из четырех приведенных. Какого рода таблицы служат и для представления результатов исследования. Эта ситуация типична для исследований общественного мнения.

Социолог называет распределение признака ***«линейкой», простым*** распределением, ***линейным*** распределением, ***частотным*** распределением, простой группировкой, потому что речь в самом деле идет о самых простых, ***одномерных распределениях*** в отличие от условных и многомерных. Последние получаются тогда, когда одновременно строится распределение по нескольким признакам. К случаю двумерных распределений перейдем чуть позже.

Одномерное распределение может быть получено как для всей выборочной совокупности, так и для отдельной подвыборки. В нашем случае подвыборкой являются студенты-гуманитарии, выделенные из всей совокупности опрошенных студентов. Тогда точнее называть распределения, полученные нами по трем признакам, ***условными.*** Такого рода условные распределения позволяют уже на этом первом этапе работы с эмпирическими данными решать задачи сравнительного анализа. Например, можно сравнивать структуру удовлетворенностью учебой студентов-гуманитариев и студентов-естественников, структуру продолжительности учебы для социологов и историков и т. д. В любом случае мы сравниваем ***структуру*** распределений для различных групп ***обследованных/опрошенных.***

Кроме такого сравнительного анализа, одномерные распределения необходимы социологу ради достижения следующих ***целей.*** *Во-первых,* для проверки качества выборки, если речь идет о массовых опросах. Даже тогда, когда выборка «хорошо»

планируется, в реальных данных могут возникнуть перекосы. Признаки, по которым формируется выборка, включаются в инструментарии, и по их распределениям осуществляется соответствующий контроль. Это только один аспект. Другой связан с тем, что число признаков, по которым планируется выборка, не может быть большим. В этой связи ряд признаков, интересующих социолога с точки зрения репрезентативности выборки, выпадают из рассмотрения при ее формировании. Тогда социолог может проверить репрезентативность по этим признакам на основе анализа их распределений.

Во-вторых, по одномерным распределениям определяется дифференцирующая сила признаков. Возвращаясь к таблице 3.1.1, видим, что две профессии не встречаются в наших данных. Соответственно, они исключаются из дальнейшего анализа. Некоторая группа (по уровню удовлетворенности, профессиональная) респондентов может быть по численности небольшой (что есть «много» и «мало», определяет социолог, исходя их своих исследовательских задач). Небольшая группа исключает возможность сравнения ее с другими, большими группами. В этом случае, опираясь на простые распределения, принимается решение и об объединении отдельных групп. Тем самым могут уточняться задачи и гипотезы исследования.

В-третьих, по простым распределениям определяем характер этого распределения и устанавливаем эмпирические закономерности «поведения» признака в отношении изучаемых объектов (в нашем случае студенты-гуманитарии). Термин «поведение» будем употреблять исключительно для наглядности и образности. На наш взгляд, он полезнее, чем математические термины.

Прежде всего, по распределениям выделяются ***модальные*** (часто встречающиеся) и ***антимодальные*** (редко встречающиеся) тенденции. Не только первые, но и вторые могут быть ***социально значимыми.*** «Мало» для социолога имеет два значения. Первое - выборка была мала по объему, и представители какой-то группы в нее не попали случайно. Второе - «редкая» группа, но социально значимая. Например, случай латентных социальных групп. Из этого вывод - нельзя выкидывать из анализа феномен «антимодальности» без достаточного обоснования.

И, наконец, представляется важным следующее. Одномерное распределение можно анализировать на разных «языках». Первый основной - язык математической статистики, статистического анализа. Огромное количество литературы описывает именно этот аспект. Основной постулат статистического подхода: одномерное распределение - результат только одного наблюдения генеральной совокупности и, соответственно, подвержено влиянию случайных, неконтролируемых, факторов. Если выборка была «хорошей», то по ней можно с определенной точностью вычислить характеристики генеральной совокупности. Отсюда и возникает понятие ***доверительного интервала,*** интервала, в котором находится истинное (для генеральной совокупности) значение такого рода характеристики. На языке статистического анализа возможные значения признака называют ***вариантами,*** а их совокупность и соответствующие им частоты - ***вариационным рядом.*** Этими терминами социологи практически не пользуются.

Второй «язык» опирается на информационный подход или понятия ***теории информации.*** Существует понятие единицы информации. Таковой является ***бит*** (от английского binary digit - двоичная цифра). Любой поток информации (числа, буквы, фразы) можно закодировать нулями и единицами. Число нулей и единиц, необходимых для оптимального (самого короткого) кодирования этого потока, называется ***количеством информации.***

Представим теперь ситуацию, когда нам надо что-то узнать. Например, кто-то из вас загадал кого-то из присутствующих. Какое число вопросов надо мне задать ему, чтобы узнать, «кого» он загадал. При этом только вопросы с вариантами ответа «да» и «нет». Для этого я составлю список из всех, например, 32 присутствующих

студентов. Затем поделю этот список на две части и спрошу, указывая на первую часть списка, «есть ли загаданный в этой части». Тем самым определю 16 студентов, среди которых есть и загаданный. Повторю процедуру деления на две части и получу список из 8-ми студентов, среди которых есть и загаданный. Продолжение такой процедуры деления приводит к результату. Мне надо было задать всего пять вопросов. Пять и есть количество информации. Это количество можно было определить и по-другому. Каждому порядковому номеру студента поставлю в соответствие пятизначное двоичное число от 00000 до 11111 и спрошу, верно ли, что у задуманного студента первая, вторая, третья, четвертая и пятая цифры равны единице?

Количество информации, необходимое для отгадывания задуманного студента, равно пяти или $\log_2 32$. В качестве упражнения подсчитайте количество информации в номере паспорта.

Одномерное распределение может интерпретироваться как некое сообщение, несущее в себе определенное ***количество информации.*** Это количество можно оценить некоторой мерой, и значение ее будет разным для разных распределений. Такая мера называется также ***энтропией.*** Если кого-то из вас заинтересует эта проблематика, то загляните в интересную книгу (10) венгерского математика, где есть раздел «Записки студента по теории информации».

Третий «язык» - просто поиск регулярности, значимость которых может описываться и без всякой математической статистики. Существуют «языки» анализа распределений, когда анализируются упорядоченности и соотношения между частотами, например, для поиска социальных констант. Но эти проблемы уже для следующего этапа изучения методологии анализа информации. «Языков» анализа распределений может быть много, поэтому это еще одна причина, по которой мы пользуемся понятием «поведение» признака, а не термином статистический анализ.

Задание на семинар или для самостоятельного выполнения

Каждому студенту необходимо придумать данные для модельной задачи. По возможности используйте фрагмент из реального исследования. Цель задания - подготовка к освоению приемов первичного анализа, т. е. изучение «поведения» отдельно взятых признаков, в том числе и эмпирических индикаторов. На этом же материале будем осваивать и анализ взаимосвязей между признаками.

Требования к задаче, а значит к эмпирическим данным, таковы:

1. Число объектов 45—50. В роли объектов могут выступать: респонденты, семьи, студенческие группы и т. д. Скорее всего, это будут респонденты, ибо объектов нужно около 50-ти. Предупреждение к «всезнайкам» - на данном этапе все делается без компьютера. Рекомендуется сначала выполнить вручную все приведенные в *«Лекциях»* задания и только потом воспользоваться компьютером.

2. Число признаков как минимум равно трем. Первый из них измерен по номинальной шкале с числом градаций, равным 6—9. Второй - по порядковой шкале с числом градаций, равным 5—7. И наконец, третий признак измерен по метрической шкале (числа, количества). При этом для упрощения вычислений в качестве значений признака рекомендуется использовать двузначные целые числа.

3. Для этих трех признаков должен иметь содержательный смысл анализ взаимосвязей между ними. Например, можно изучить «поведение» таких признаков, как «социальное происхождение студента», «его уверенность в трудоустройстве по специальности после окончания вуза» и «отношение к учебе». При этом первый из них имеет номинальный уровень измерения и представляет собой прямой вопрос анкеты о социальном происхождении. Второй может быть измерен посредством логического квадрата по пятибалльной порядковой шкале. Третий измерен по шкале Терстоуна и тем самым имеет метрический уровень измерения.

4. Для выбранных признаков должны быть правомерны, например, такие вопросы: «Зависит ли уверенность в трудоустройстве от социального происхождения студента?», «Зависит ли отношение к учебе от уверенности в трудоустройстве?».

5. После выбора исходных для анализа признаков следует сочинить ответы, если задача модельная. Таким образом, получается матрица исходных данных вида «объект - признак», на основе которой будут выполняться задания к нескольким последующим разделам этой главы.

6. По всем трем признакам необходимо вычислить абсолютные, относительные (в долях и процентах) и накопленные частоты. Оформить результаты в виде таблиц типа 3.1.1; 3.1.2; 3.1.3.

2. АНАЛИЗ ХАРАКТЕРА «ПОВЕДЕНИЯ» ПРИЗНАКА

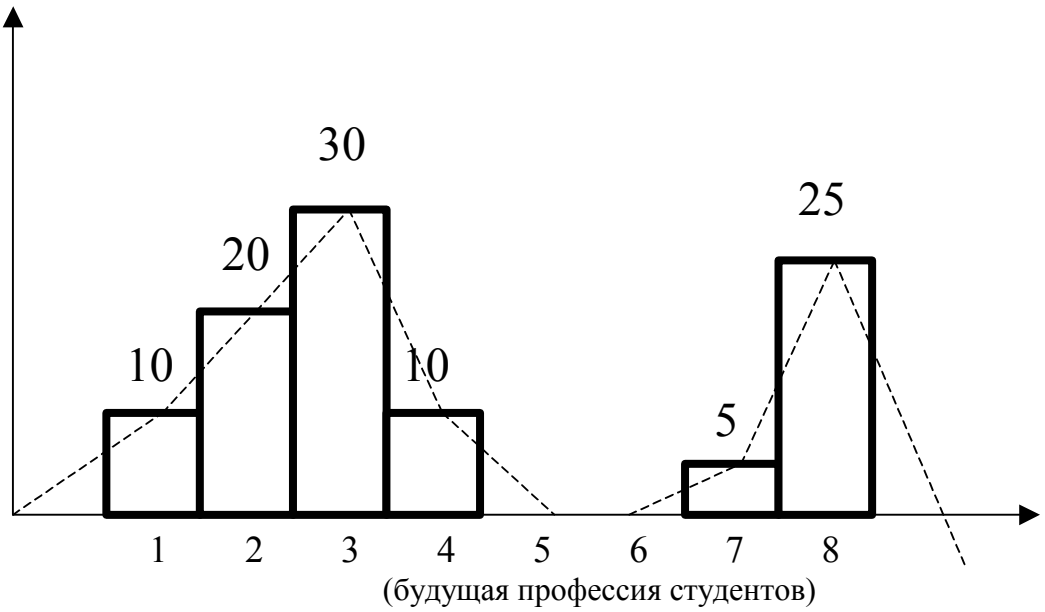
Эмпирическая кривая распределения. Показатели средней тенденции для различных типов шкал. Дескриптивная статистика. Мода. Медиана. Среднее арифметическое значение, взвешенное среднее. Меры рассеяния вокруг средних. Дисперсия. Коэффициент вариации как мера однородности. Квартильный размах. Меры качественной вариации. Коэффициент качественной вариации. Среднее геометрическое. Энтропия.

Регулярно на экране телевизора вы видите ***визуально*** представленные ***распределения*** какого-нибудь признака (столбики с обозначением процентов). Например, результаты изучения общественного мнения по претендентам на президентский пост или место в парламенте. Эти картинки называются ***гистограммами*** - графическое изображение или визуализация распределений. Они строятся по определенным правилам и в основном нужны не столько самому социологу, сколько заказчику социологического исследования (красиво и наглядно). Социологу они нужны лишь на предварительном этапе работы с эмпирией для того, чтобы на компьютере быстро просмотреть характер распределений. Существует множество способов визуализации. Например, в работе [2] приводится 15 способов визуального изображения (графики, диаграммы) одних и тех же данных - одномерного распределения признака.

На рис. 3.2.1 изображена гистограмма, соответствующая распределению студентов по будущим профессиям. На горизонтальной оси, начиная с любой точки, откладываются на равном расстоянии восемь (см. таблицу 3.1.1) профессий. Над каждой «профессией» воздвигается столбик высотой равный относительной частоте этой профессии. Столбики могут отстоять друг от друга и на каком-то расстоянии. В нашем случае они примыкают друг к другу. Гистограмму можно строить по частостям или по процентам. Они совпадут при соответствующем выборе масштаба. Для этого на вертикальной оси одна и та же точка должна соответствовать либо единице, либо ста процентам.

Процент/ частость/

Рис. 3.2.1 Гистограмма и эмпирическая кривая распределения студентов по профессиональным группам



Сумма площадей всех прямоугольников равна единице, если она построена по частостям и равна ста, если гистограмма построена по процентам. Вертикальная ось служит только для задания масштаба, поэтому гистограмму начинают строить с любой позиции по горизонтали. Ломаная линия (обозначенная на рис. 3.2.1 пунктиром) называется ***эмпирической кривой распределения,*** или ***полигоном.*** Она соединяет середины верхней стороны прямоугольников. Эта кривая и ее характеристики говорят социологу о «поведении» признака. Второй из этих терминов мало употребляется на практике.

Аналогичным образом строится гистограмма и эмпирическая кривая распределения для второго признака, т. е. для распределения студентов по степени их удовлетворенности учебой. Они изображены на рис. 3.2.2. Если для номинальных и порядковых шкал гистограммы эмпирическая кривая распределения служит только для визуализации, то для метрических они имеют особый смысл.

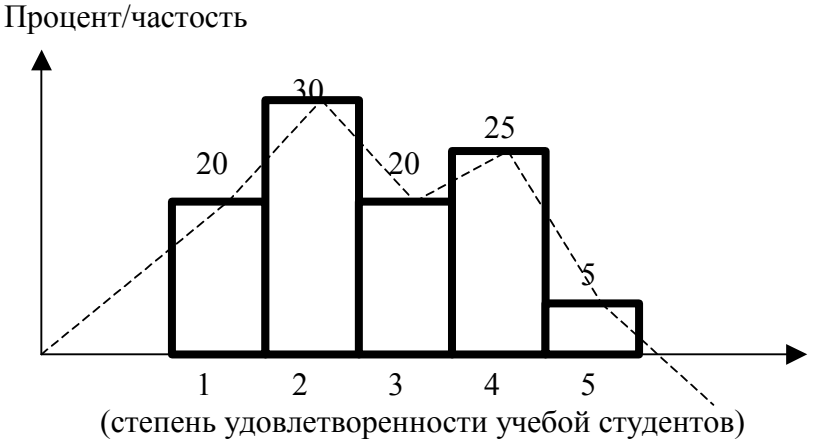


Рис. 3.2.2 Гистограмма и эмпирическая кривая распределения по степени удовлетворенности учебой

Построим гистограмму и эмпирическую кривую распределения для признака «продолжительность затрат времени на учебу». В этом случае гистограмма строится несколько иначе. Как вы заметили, каждый столбик гистограммы по площади был равен числу респондентов. Визуально передается не высота столбика, а его площадь.

Ширина столбика равнялась единице и для номинального, и для порядкового признаков. В данном случае ширину нельзя выбрать одинаковой, так как наши интервалы разные. Поэтому гистограмма строится по ***плотности*** распределения. Плотность в интервале -это число респондентов, приходящихся на единицу интервала. Обозначим плотность в наших шести интервалах через

$\rho_1, \rho_2, \rho_3, \rho_4, \rho_5, \rho_6$

Тогда $\rho_1 = 27/1 = 27$; $\rho_2 = 75/1.5 = 50$; $\rho_3 = 150/1.5 = 100$;
$\rho_4 = 348/3 = 116$; $\rho_5 = 250/1 = 250$; $\rho_6 = 150/1 = 150$

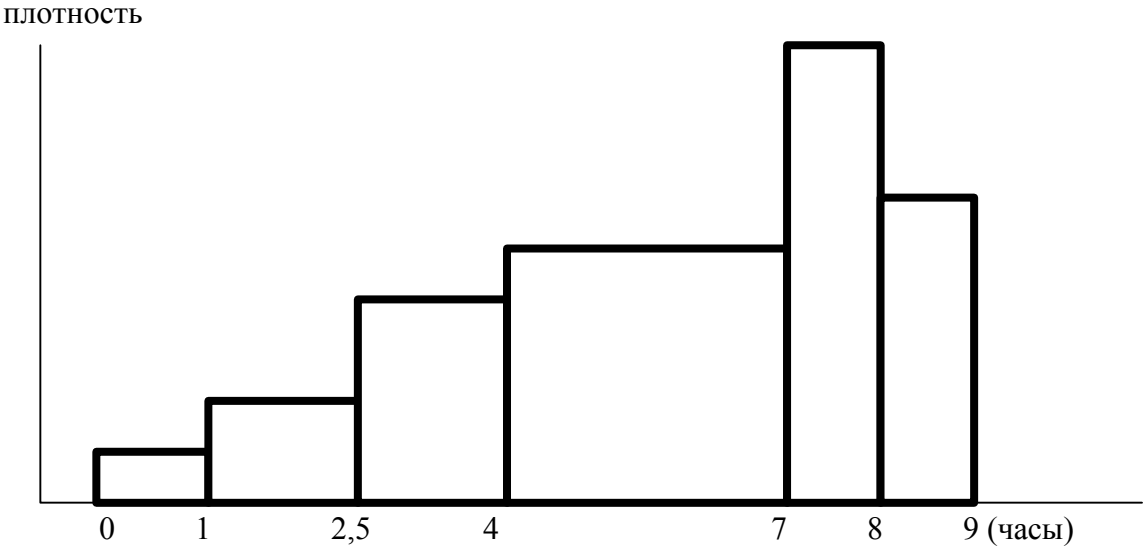


Рис. 3.2.3 Гистограмма по продолжительности затрат времени на учебу

В данном случае эмпирическая кривая распределения не имеет содержательного смысла, ибо не передает характера распределения. Поэтому такую кривую строят при делении на равные интервалы. Число интервалов при этом определяется уже исходя из формальных критериев. Для порядковой и метрической шкалы гистограмму и эмпирическую кривую распределения можно построить и по накопленной частоте. Только в этом случае для эмпирической кривой распределения существует специфическое название. Она называется ***кумулята,*** а накопленную частоту называют кумулятивной. Построим ее по данным, представленным в таблице 3.2.1.

Таблица 3.2.1

Распределение по продолжительности учебы (равные интервалы)

Показатели	**Продолжительность учебы**									**Итого**
	0-1	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**	**8-9**	
Абсолютная частота	27	50	75	100	48	100	200	250	150	1000
Относительная частота в %	2,7	5,0	7,5	10	4,8	10	20	25	15	100
Накопленная частота	2,7	7,7	15,2	25,2	30	40	60	85	100	

На рис. 3.2.4 изображены гистограмма и кумулята по продолжительности затрат времени на учебу (интервалы равные, их девять). Кумулята - это всегда возрас-

тающая кривая. Пока на пунктирные линии не обращайте внимания.

Графическое изображение распределений в виде эмпирических кривых распределения (полигоны и кумуляты) нужны социологу в зависимости от типа шкал для разных целей. Для номинальной шкалы мы можем упорядочить (провести ранжирование) различные профессиональные группы по их представительности (объему) в наших данных и соответственно выделить ***модальные*** (самые большие по объему) группы. Для порядковой шкалы, кроме этого, определяется и степень единодушия студентов в оценке своей удовлетворенности учебой. Вспоминаем шкалу Терстоуна, для построения которой посредством ***медианы*** **и** ***квартального размаха*** оценивалась степень единодушия экспертов. Самую важную роль играют эмпирические кривые распределения для метрических признаков. Но эта роль связана не с первичным анализом и не с изучением поведения эмпирических индикаторов, а с анализом поведения показателей/коэффициентов/ индексов.

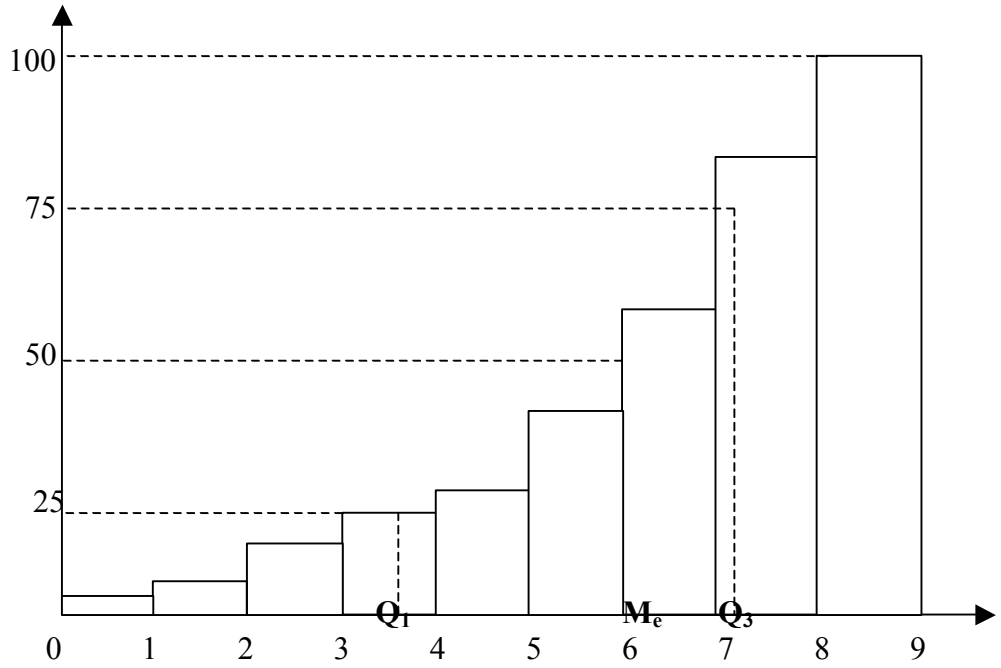


Рис. 3.2.4 Кумулята по продолжительности затрат времени на учебу

При статистическом подходе к анализу распределений каждый такой показатель теоретически может иметь ***закон распределения*** *с определенными параметрами и по эмпирической кривой распределения можно судить о том, каков этот закон. Знание законов дает возможность применения к анализу эмпирии всего богатства средств, накопленных в математической статистике. Законов очень много, и отсюда названия: нормальный закон распределения (рис. 3.2.5), логарифмический закон распределения (рис. 3.2.6), линейный закон распределения (рис. 3.2.7) и т.д. Законы вы проходили и в школе. Уравнение прямой, параболы, гиперболы интерпретируются как математические законы, связывающие две величины X и Y. Некоторые законы нельзя записать в явном виде, т. е. в виде математической формулы.*

Что касается самого факта существования закона распределения какого-то показателя, то это требует доказательства. Например, в виде проверки статистических гипотез. Эту тему относим к последующим этапам в вашем образовании.

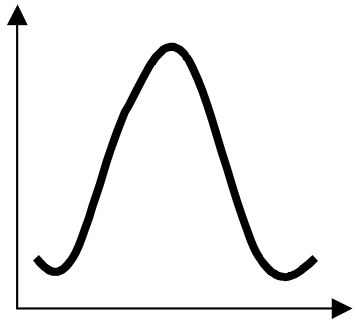


Рис. 3.2.5

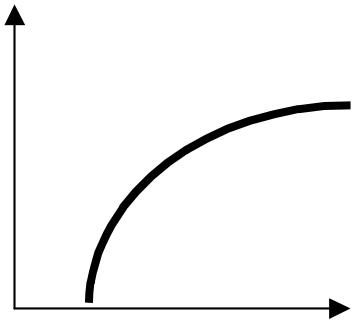


Рис. 3.2.6

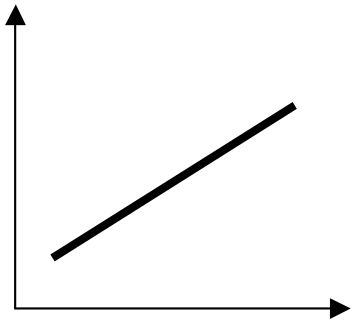


Рис. 3.2.7

Перейдем к рассмотрению характеристик, описывающих (отсюда название дескриптивная статистика) «поведение» признака в целом, в виде некоторой эмпирической тенденции. Потому они и называются мерами центральной тенденции.

Мода

Наиболее часто встречающееся значение признака называется модой. Таких значений может быть и несколько. В нашем случае третья профессия является модальной. Социолог никогда не работает с одной единственной модой, а употребляет понятие «модальные значения». Для нашего примера профессии 3 и 8 являются модальными. Аналогична ситуация в случае порядковых шкал. Мода равна 2 (наиболее часто встречаются студенты, степень удовлетворенности учебой которых равен двум). В качестве модальных значений имеет смысл рассматривать все же два значения, 2 и 4, т. е. наиболее распространены две группы по степени удовлетворенности. И это несмотря на то, что по объему они различны. Однако по сравнению с другими группами они достаточно большие. Можно считать, что наличие таких модальных групп специфично, характерно, типично для изучаемой совокупности студентов-гуманитариев. Это самая простая эмпирическая закономерность.

Нахождение модального значения в случае метрической шкалы невозможно по рис. 3.2.3, ибо ширина интервалов различна и это модальное значение может находиться в любом интервале. Поэтому, прежде всего, возникает задача определения ***модального интервала*** - интервала, содержащего моду. Для этого необходимо перейти от деления на интервалы, основанного на содержательных критериях, к делению на интервалы по формальным критериям. При этом интервалы должны иметь равную длину и их число должно зависеть от степени изменчивости признака. Чем больше степень изменчивости, тем больше нужно интервалов для определения модального. На рис. 3.2.8 приведена гистограмма, построенная для случая деления «продолжительности» на девять равных интервалов. Абсолютные частоты в этих интервалах были приведены выше в таблице 3.2.1. Плотность в каждом интервале пропорциональна этим абсолютным частотам. Ширина интервала равна 1. Эмпирическая кривая распределения в этом случае называется эмпирической функцией распределения плотности.

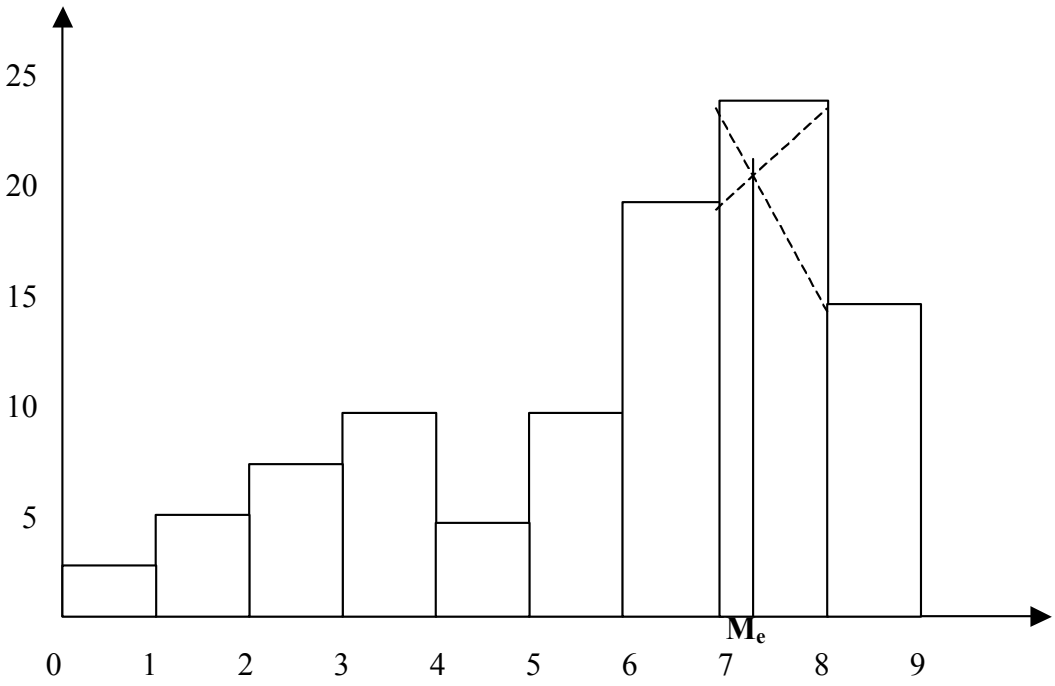


Рис. 3.2.8 Гистограмма «продолжительность учебы» (равные интервалы)

Существует математическая формула для вычисления моды, но мы приведем лишь геометрический способ нахождения моды в модальном интервале. Модальным интервалом является интервал в 7—8 часов. Значение моды вычисляется геометрически (пересечение пунктирных линий на рис. 3.2.8) и примерно равно 7,3 часа (см. стрелочку на том же рисунке). Является логичным, что мода должна находиться ближе к тому концу модального интервала, который примыкает к интервалу с большим числом объектов. Возникает вопрос, как подсчитать значение моды, если модальный интервал первый или последний по счету. Тогда за моду принимается середина этих интервалов.

Модальные значения определенным образом говорят о характере поведения признака и в основном о числе «горбов». Например, вспоминаем задачу ранжирования по предпочтениям различных сортов пива. С какими ситуациями мы сталкивались? С достаточным единодушием (один горбик, одна мода), с двумя противоположными тенденциями (два горбика, две моды) и с полным разнообразием (практически равномерное распределение - моды нет). Чтобы как-то продвинуться в анализе предпочтений, мы использовали еще одну характеристику - медиану, к рассмотрению которой и переходим.

Медиана

Эта мера центральной тенденции, или характеристика распределения, имеет смысл только для порядковых и метрических шкал. С медианой мы сталкивались при построении шкалы Терстоуна и опять же в процедуре ранжирования. В общем случае медиана - значение признака, соответствующее середине упорядоченного ряда. Например, пусть у нас есть данные по каждой области - доли голосов в %, отданных избирателями на выборах господину Икс. Тогда значение медианы, равное 15%, интерпретируется следующим образом. В половине областей отдано за господина Икс больше 15% голосов, а в половине - меньше 15%. Не правда ли, это очень важная характеристика для интерпретации результатов выборов?

Для вычисления медианы в этом случае мы должны были упорядочить все области в порядке возрастания или убывания числа голосов. Если число областей не-

четное, то в середине ряда - одна единственная область. Медиана тогда равна числу голосов, отданных господину Икс в этой области. Если число областей четное, то середину ряда составляют две области и медиана вычисляется как среднее значение по этим двум областям.

В случае нашего примера метрической шкалы - продолжительность затрат времени на учебу - медиана может быть вычислена таким же образом. Для этого проведем упорядочение студентов по возрастанию/убыванию этих затрат и найдем середину аналогичным образом. Медиану можно вычислить и по кумуляте (см. шкалу Терстроуна).

Для порядковых и метрических шкал необходимым является понятие ***медианного интервала,*** т.е. интервала содержащего медиану. Как правило, вы не любите формулы, поэтому приведем вербальное описание формулы для вычисления медианы в медианном интервале. Это делается по двум соображениям. Первое - показать, что математическая формула всегда отражает содержание. Второе - математической формулой иногда пользоваться удобнее для избежания очень длинных описаний. Итак, медиана в медианном интервале вычисляется по формуле:

Me = (нижняя граница медианного интервала) (x) + (ширина медианного интервала) (l) x [(половина от числа объектов) (n/2) - (частота, накопленная до медианного интервала) (P)] : (частота в медианном интервале) (p)

Эту формулу можно записать очень просто с использованием обозначений, приведенных внизу:

$$Me = x + l\frac{\frac{n}{2} - P}{p}$$

Чем выше уровень измерения, тем богаче возможности описания «поведения» признака. Если признак измерен по метрической шкале, то кроме моды и медианы для описания поведения признака используется известная всем мера центральной тенденции - средняя арифметическая.

Среднее арифметическое

Для любой совокупности значений признака это сумма всех значений, деленная на их число. Вернемся к примеру признака - продолжительность затрат времени на учебу. Обозначим число студентов-гуманитариев через n (для нашего случая n = 1000), а через X_i - значение этой продолжительности для 1-го студента. Тогда средняя арифметическая продолжительности будет равна:

$$X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Таким образом можно определить среднюю продолжительность затрат времени на учебу в группах студентов с любой «будущей профессией», с любой степенью удовлетворенности учебой и т. д.

Социолог часто встречается с ситуацией, когда конкретные значения признака по отдельным объектам неизвестны. Исходно имеются только интервалы изменения признака и частота (абсолютная или относительная) встречаемости объектов в этих интервалах. Например, та же продолжительность может быть задана в виде интервалов и частоты в них. Это может быть в двух случаях. Первый - данные о продолжительности получены с помощью прямого вопроса анкеты: «Сколько времени Вы в среднем в неделю тратите на занятия, связанные с учебой?». При этом предлагаются заданные заранее интервалы. По сути, мы имеем дело с порядковой шкалой. В этом

случае также можно вычислить среднее значение продолжительности для некоторой группы студентов. Только она называется ***средняя взвешенная*** и вычисляется несколько по-другому.

Второй случай, когда у социолога отсутствуют конкретные значения по каждому объекту в ситуации вторичного анализа. ***Вторичным анализом*** социолог называет анализ «чужих» данных для решения своих собственных, новых задач. Тогда часто приходится работать с уже вычисленными до него средними арифметическими. Например, результаты исследования бюджетов времени обычно публикуются в виде средних затрат времени с указанием объема группы, для которой они получены. В процессе вторичного анализа возникает необходимость объединения каких-то групп и, соответственно, в подсчете общей средней. В этой ситуации также необходима средняя взвешенная для вычисления «средней средних».

Вычислим среднюю продолжительность затрат времени на учебу студентами-гуманитариями по данным таблицы 3.1.3. .Для этого предполагается, что продолжительность для каждого респондента, отнесенного к интервалу, равна середине интервала. Для наших шести интервалов их середины соответственно равны:

x_1=0,5; x_2= 1,75; x_3=3,25; x_4=5,5; x_5=7,5; x_6=8,5.

Нам известно число студентов в каждом интервале:

n_1=27; n_2=75; n_3=150; n_4=348; n_5=250; n_6=150.

Тогда продолжительность затрат времени на учебу в среднем на студента или средняя взвешенная продолжительность равна:

X=(0,5×27+1,75×75+3,25×150+5,5×348+7,5×250+8,5×150)/1000=5,7

Формула для вычисления средней взвешенной выглядит для k интервалов следующим образом:

$$X = \frac{\sum_{j=1}^{k} n_j x_j}{\sum_{j=1}^{k} n_j}$$

, где x_j – середина j-го интервала.

Аналогично вычисляется «средняя средних». Допустим, перед социологом стоит задача вычисления средней продолжительности жизни мужчин в России по данным отдельных областей. Эти данные представляют собой среднюю продолжительность жизни мужчин по каждой области. Естественно, «среднюю средних» вычисляем с весами, равными численности мужчин в каждой области.

Все рассмотренные характеристики: мода, медиана, средняя арифметическая, среднее взвешенное - являются ***средними.*** Они характеризуют ***центральные*** тенденции одномерного распределения. Есть и другие средние, но они в социологии применяются редко. Поэтому среднюю арифметическую называют просто средней, а мода и медиана сохраняют свои названия. Без процедуры усреднения социолог-эмпирик существовать не может. Другое дело, с помощью каких средних он проводит эту процедуру.

Сами по себе значения «средних» мало о чем говорят, если социолог не видит эмпирическую кривую распределения, например, на экране компьютера. В ситуации «невидения» ему помогают интерпретировать любые средние так называемые ***меры вариации, меры рассеяния*** объектов вокруг этих средних. Сначала мы рассмотрим меру вариации для случая метрической, шкалы, а затем – для порядковой и номинальной.

Прежде чем перейти к этой проблеме, заметим, что любая средняя характеризует центральную тенденцию распределения только тогда, когда объекты в основном сосредоточены вокруг этих средних, т.е. изучаемая совокупность объектов однородна относительно признака. ***Однородность*** *- это очень важное понятие для всех, кто работает с эмпирией. Социолог сталкивается с проблемой однородности*

в разных контекстах. Как раз вот здесь пара понятий «качество - количество» очень важна. Разделение понятий ***качественная однородность и количественная однородность*** *имеет огромный смысл. Например, разве есть смысл в среднем доходе или в среднем возрасте россиянина? Конечно же, нет. И в то же время есть смысл в средней заработной плате сельских врачей или в среднем возрасте мужчин-пенсионеров. Необходима качественная однородность для того, чтобы начать анализ количественных характеристик распределения признака.*

Сами количественные характеристики могут указывать/показывать на отсутствие количественной однородности по анализируемому признаку. Это в свою очередь будет говорить о наличии качественной неоднородности.

Дисперсия

Рассмотрим меру вариации/рассеяния/разброса/изменчивости для метрической шкалы. По эмпирической кривой распределения или гистограмме на рис. 3.2.3 видим, что совокупность студентов неоднородна по продолжительности затрат времени на учебу. *С* одной стороны, очевидно, что средняя продолжительность учебы как характеристика имеет смысл, поскольку вполне правомерно сравнение средней продолжительности учебы для выделенных нами групп студентов: социологов, политологов, культурологов и т. д. С другой стороны, в ситуации неоднородности такое сравнение содержательно ни о чем не говорит.

Какова может быть мера неоднородности/однородности по продолжительности? Об этом можно судить по степени отклонения продолжительности затрат времени на учебу отдельного студента от средней продолжительности, которая в нашем случае равна 5,7 (в часах). Индивидуальные отклонения (x_i — X) нельзя просто суммировать, чтобы судить об общем отклонении. Отклонения в одну сторону будут погашаться отклонениями в другую. Чтобы этого не было, индивидуальные отклонения возводятся в квадрат, а затем складываются. Эта сумма делится на число респондентов, и получается характеристика, называемая ***дисперсией*** (σ^2). Это мера вариации значений признака в среднем и вокруг средней арифметической.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - X)^2}{n}$$

Следует заметить, что при небольшом числе объектов делить нужно не на n, а на (n - 1). Для социолога это не принципиально, так как он работает обычно с достаточно большим числом объектов.

Корень квадратный из дисперсии называется ***среднеквадратическим отклонением*** *(*σ-сигма). По ней можно сравнивать меры рассеяния разных признаков, одного признака для различных совокупностей. Прямое сравнение дисперсий, среднеквадратических отклонений мало что дает. Рассмотрим пример из нашего исследования. Вычислим среднее арифметическое и среднеквадратическое отклонение продолжительности затрат времени на учебу для нескольких групп студентов. Допустим, что для социологов (X=6,σ=4), психологов (X=5,4, σ=3,5), политологов (X=4,5, σ=3,5), историков (X=6, σ=2). Какие выводы можно сделать по этим данным?

Социологи и историки затрачивают на учебу в среднем одинаковое время, но совокупность социологов менее однородна, потому что среднеквадратическое отклонение больше. Психологи затрачивают на учебу в среднем больше времени, чем политологи, и они более однородны, чем группа политологов. Дисперсия одинакова в этих группах, относительно разных по значению средних. Когда средние и дисперсии в сравниваемых группах различны, на помощь приходит коэффициент вариации.

Коэффициент вариации

Этот коэффициент при наших обозначениях равен $V = \frac{\sigma}{X} \times 100$.

Он представляет собой долю вариации в процентах (%), приходящуюся на единицу средней. В нашем случае соответственно четырем группам: V1= 66,7% (для социологов), V2=64,8% (для психологов), V3=77,8% (для политологов), V4=33,3% (для историков). Таким образом, группа историков более однородна по продолжительности затрат времени на учебу, чем все остальные группы. Самая неоднородная группа - политологи. Это означает, что среди них оказались и очень много, и очень мало занимающиеся.

Среднее арифметическое и дисперсия интерпретируются всегда вместе. Например, существует так называемое правило *«трех сигм»,* очень важное при работе с эмпирией. Оно означает, что если все значения признака находятся в интервале от -3σ до +3σ, то считается, что закон распределения признака ***нормальный,*** т. е., как минимум, эмпирическая кривая имеет унимодальный характер (одна мода, один горб). На рис. 3.2.5 изображен идеальный нормальный закон распределения. Запомните его, ибо математический аппарат для анализа нормальных распределений очень богат. Для идеально нормального распределения мода, медиана и среднее арифметическое равны.

Если для анализа распределений использовать «язык» статистического анализа, то сами рассмотренные характеристики, например X, являются величинами, имеющими свой собственный закон распределения. Представим себе, что каждый из вас для одного и того же исследования сформировал выборочную совокупность. Пусть у каждого будет самая из самых «хорошая» (репрезентативная) выборка. Если подсчитать, к примеру, средний возраст опрошенных по этим выборкам, то значения будут различны. Среднее этих значений и будет истинным значением среднего возраста в генеральной совокупности. Аналогичны рассуждения и в случае средней продолжительности затрат времени на учебу.

Отклонение средних от «истинной средней» будет носить случайный характер. Оказывается, эту случайность можно оценить. На этом основан подсчет так называемых доверительных интервалов, т. е. интервалов, в которых находится истинное (для генеральной совокупности) значение признака. Но это только для тех величин (характеристик), для которых известен закон распределения. Они называются **статистиками.** Среднее арифметическое и является статистикой с нормальным законом распределения. Для нее легко определяется доверительный интервал.

Другие меры вариации

Рассмотрим меру вариации, меру отклонения, меру рассеяния значений признака вокруг медианы. Такой мерой является ***квартильный размах,*** *с* которым мы встречались при построении шкалы Л. Терстоуна. Вспомним, что содержательно это интервал, в котором вокруг медианы сосредоточилось 50% экспертов. Это единственная мера вариации для порядковых шкал. На рис. 3.2.4 три пунктирные линии проведены для определения медианы и соответствующего ей квартильного размаха {он равен $(Q_3-Q_1)/2$}. Без сравнительного контекста трудно сказать, мало это или много. Для социолога ***познавательная возможность*** любого математического конструкта, а это пока простейшие формулы на уровне обыденного понимания, определяются только в сравнительном контексте, т. е. при сравнении значений, полученных в разных условиях.

Перейдем к самым трудным для понимания мерам - мерам качественной вариации, т. е. мерам вариации для признаков, измеренных по номинальным шкалам. Самое главное, что любая такая мера характеризует степень отклонения распределения признака от ***равномерного,*** т. е. когда каждой градации признака соответствует одно

и то же число объектов. Максимальное значение меры обычно соответствует ситуации равномерного распределения, а минимальное - ситуации, когда все объекты сосредоточены в одной градации.

Как мы знаем, любой номинальный признак сводится к совокупности бинарных, дихотомических, т. е. принимающих значения 0 или 1. В этом случае столбец нашей исходной матрицы данных «объект-признак», соответствующий одному признаку, превращается как бы в несколько столбцов, каждый из которых соответствует отдельному свойству (быть социологом, быть политологом и т. д.). Анализировать мы должны теперь поведение «свойства», а не признака. По всем объектам это совокупность из нулей и единиц.

0 0 0 0 1 1 1 1 1 1 ... 0 0 1 1 1

Предположим, что этот ряд получен по свойству - быть в будущем социологом. Если i-й студент - социолог, то ему соответствует $x_i=1$, а если он не социолог, то $x_i=0$. Оказывается, для такого вида данных имеет смысл среднее арифметическое. Она равна X=k/n, где k - число будущих социологов, а n- число всех студентов-гуманитариев.

Почему имеет смысл средняя арифметическая для дихотомической шкалы? Потому что она содержательно интерпретируется. Если X=0, то это означает, что все студенты-гуманитарии в нашей выборке не социологи. Если X=l, то все студенты - социологи. Если X=0,5, то половина студентов - будущие социологи, а половина - не социологи. Продолжая наши рассуждения, можно сделать вывод и для случаев, когда $0 < X < 0{,}5$ и $0{,}5 < X < 1$. Первый из них означает, что в совокупности меньше 50% студентов социологи. Второй - в совокупности больше 50% социологов.

Таким образом, как это ни парадоксально, можно вычислять среднее арифметическое по признаку «пол». Только важно правильно интерпретировать полученный результат, исходя из того, каким образом закодирован этот признак. Разумеется, социологу нет никакого смысла в использовании такого рода средней, отражающей «центральную тенденцию». Он прекрасно работает с относительными частотами в %. Приведенная средняя интересна не для целей первичного анализа, а для анализа с применением сложных математических методов. К примеру, для такой средней можно подсчитать дисперсию. Если для дихотомических признаков имеет смысл использование характеристик метрической шкалы, значит, возможно использование и математических методов, работающих с метрическими данными. Дисперсия в данном случае равна:

$\sigma^2 = \frac{k(n-k)}{n^2}$, т.к. в формулу

$\sigma_i^2 = \frac{\sum (x_i - x)^2}{n}$ нужно подставить

$x = \frac{k}{n}$ и учесть, что k раз встречается 1 и (n-k) раз 0, поэтому

$$\sigma^2 = \frac{(n-k)(0-\frac{k}{n})^2 + k(1-\frac{k}{n})^2}{n} = \frac{(n-k)k^2 + k(n-k)^2}{n^3} = \frac{k(n-k)(k+n-k)}{n^3} = \frac{k(n-k)n}{n^3} = \frac{k(n-k)}{n^2}$$

Эта дисперсия и является мерой вариации для бинарного (дихотомического) признака. При этом она равна нулю, если все объекты либо обладают, либо не обладают анализируемым свойством. Что естественно, так как в этих случаях разброса в данных не наблюдается. Максимальное значение этой дисперсии достигается в случае равномерного распределения (k=n/2), и оно равно 1/4. При этом $X=1/2$, $\sigma=1/2$, V=100%.

Напомню вам одно правило из школьной арифметики. Если есть два целых числа, то среднее геометрическое этих чисел всегда меньше или равно среднему арифметическому. Равенство достигается, когда числа равны.

$$\sqrt{ab} \le (a+b)/2$$

Этим соотношением и воспользуемся для введения коэффициента качественной вариации. Вначале предположим, что номинальный признак имеет только две градации, причем в первую градацию попало N_1 объектов, а во вторую – N_2 объектов (число всех объектов равно $n=N_1+N_2$). И если теперь в соотношение между средней арифметической и средней геометрической подставить

$a = N_1^2$; $b = N_2^2$, то получим $N_1 \times N_2 \le (N_1^2 + N_2^2)/2$

Максимальное значение $N_1 \times N_2$ будет только в случае $N_1 = N_2$, и оно будет равно $n^2 / 4$. А это ведь случай равномерного распределения. ***Коэффициентам качественной вариации*** и будет отношение реального значения произведения ($N_1 \times N_2$) к максимальному его значению, равному $n^2 / 4$.

Коэффициент равен нулю, если все объекты в одной градации, и единице, если распределение равномерное. Коэффициент легко обобщается на случай, когда число градаций равно k. Представим себе, что из всей совокупности объектов мы образовали всевозможные пары. Вспомним метод парных сравнений Терстоуна и вычисление числа всевозможных пар для сравнения объектов. Здесь ситуация аналогичная. Пары не повторяются, объект сам с собой пару не образует. В случае двух градаций произведение ($N_1 \times N_2$) есть не что иное, как число пар, различных между собой.

Если градаций три и по ним частоты равны (N_1, N_2, N_3), то число различных пар будет равно ($N_1 \times N_2 + N_1 \times N_3 + N_2 \times N_3$). Число членов в этой сумме вычисляется как число парных сочетаний из трех элементов по два. Вспоминаем, что это число равно k(k-l)/2, когда число элементов равно k.

Тогда коэффициент вариации вычисляется как ***отношение:***

- реального числа различных пар, равного ($N_1 \times N_2 + N_1 \times N_3 + N_2 \times N_3$);
- к максимальному (случай равномерного распределения), равному $\{(n^2/9)(3 \times 2/2)\}$. В первых круглых скобках - то, во что превращается каждый член суммы, а во вторых - число членов в этой сумме.

В общем случае для k градаций реальное число пар равно $\{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_i N_j\}$, а максимальное - $\{(n^2/k^2)(k(k-l)/2)\}$. Таким образом, формула для вычисления коэффициента качественной вариации приведена по частям, т. е. отдельно числитель (реальное) и отдельно знаменатель (максимальное).

Коэффициентом вариации (R) может служить и величина, равная *среднему* ***геометрическому*** из относительных частот в долях (частости) умноженному на число градаций, т. е.

$$R = k\left(\sqrt[k]{\frac{N_1}{n}\frac{N_2}{n}\cdots\frac{N_k}{n}}\right)$$

Для вычисления этой величины необходимо избавиться от пустых градаций, иначе она обратится в нуль. R=l при равномерном распределении.

Приведем еще один пример вычисления меры качественной вариации. В качестве такой меры служит ***энтропия,*** о которой мы упоминали в контексте «языка» анализа распределений, опирающегося на информационный подход. Энтропия - это основное понятие теории информации. Распределение признака интерпретируется как некое сообщение, несущее определенный объем информации. Этот объем можно

оценить энтропией как мерой «определенности»/«неопределенности». Ее трудно объяснить и трудно понять без знания логарифмов и логарифмических законов распределения. Более того, замечательные свойства этой меры могут быть оценены только при многомерном анализе. Пока вам придется просто этому поверить. Итак, энтропия H(x) при числе градаций равном k и при обозначении j-й частости (доли) через p_i равна:

$$H(x) = -\sum_{i=1}^{k} p_i \log p_i$$

Логарифм может быть взят по любому основанию, ибо нетрудно перейти от одного основания к другому. Напомним, что есть натуральный логарифм (по основанию «e»), десятичный (по основанию «10»), двоичный (по основанию «2»).

Энтропия - положительная величина, несмотря на то, что перед суммой стоит минус. Он погашается другим минусом, появляющимся за счет того, что логарифм берется от правильной дроби (это вам известно из школьной математики). Значение энтропии равно нулю, если все объекты сосредоточены в одной градации (но чтобы это показать, нужны знания о «пределах» - lim). В самом деле, тогда мера ***неопределенности*** минимальная. Энтропия равна log k, если распределение равномерное, т. е. в этом случае максимальная неопределенность. Чтобы значение меры не зависело от числа градаций, можно использовать в качестве меры качественной вариации нормированную величину энтропии.

Термин нормировка будет дальше встречаться часто. Это процедура преобразования некоторой величины в необходимый для исследователя вид. Она нужна для того, чтобы какие-то показатели/коэффициенты/ индексы изменялись либо от 0 до 1, либо от -1 до +1. Тогда делается возможным сравнение их значений, полученных при разных условиях, например, для различных совокупностей объектов.

На практике пользуются в сравнительном контексте только одной мерой качественной вариации, ибо каждая мера отражает свое собственное понимание вариации. Потому значения, полученные по разным мерам, не имеет смысла сравнивать.

Анализ «поведения» динамических рядов

Коротко остановимся на анализе динамических рядов. Эмпирическая кривая распределения в этом случае строится по конкретным значениям признака. На рис. 3.2.9 изображен динамический ряд - изменение коэффициента рождаемости за сто лет в некоторой стране X. По горизонтали обозначены 10 точек, каждая из которых соответствует пятилетнему интервалу. По вертикали отложены значения коэффициента рождаемости в среднем за соответствующую пятилетку. Пример модельный. Мы не знаем, какая это страна, и какое это столетие.

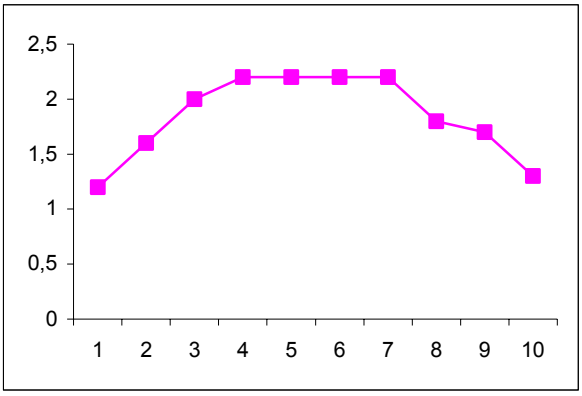


Рис. 3.2.9 Динамический ряд изменения рождаемости

Все рассмотренные выше меры центральной тенденции могут использоваться и для анализа временных рядов. Если изменения значения признака наблюдаются (как

в нашем случае), то основным вопросом при анализе временных рядов является его «выравнивание» и определение ***«тренда»,*** т. е. кривой, характеризующей общую тенденцию изменения признака, т. е. закон поведения коэффициента рождаемости. Другими словами, появляется необходимость в описании эмпирической кривой с помощью математической функции или определение теоретического закона распределения, максимально приближенного к эмпирической кривой. Только после определения тренда можно предсказать значение признака в следующих временных точках. Кстати сказать, найти закон не всегда удается. Тогда анализ проводится по отдельным частям эмпирической кривой распределения.

Если на эмпирической кривой распределения наблюдаются цикличности, то выравнивание заменяется сглаживанием «скользящей средней» из значений, число которых охватывает цикл. Можно изучать и ***«лаги».*** «Лаг» - показатель опережения или отставания одного явления (в нашем случае коэффициента рождаемости) от другого. Например, от мероприятий, принятых для повышения рождаемости.

Существует целая область науки, которая занимается проблемами анализа динамических рядов. В социологии такие ряды встречаются при работе с первым из пяти выделенных нами типов информации, а именно с государственной статистикой. В основном с временными рядами работают специалисты в области анализа социальных систем и социальной демографии.

Задание на семинар или для самостоятельного выполнения

Задание выполняется индивидуально и состоит из следующих этапов:

1. По данным первых двух таблиц, полученных каждым студентом в рамках предыдущего задания, необходимо построить гистограммы. Убедиться в том, что гистограммы построенные для признака по абсолютным частотам, долям и процентам, будут совпадать при выборе определенного масштаба.
2. Подсчитать для третьего признака плотность в каждом интервале. Построить гистограмму по плотности.
3. Изобразить на гистограммах эмпирическую кривую распределения.
4. Построить по накопленной частоте гистограмму для порядковой шкалы и изобразить кумуляту и геометрически определить медиану в медианном интервале. Геометрически определить квартильный размах.
5. Разбить метрическую шкалу на равные интервалы (порядка 15-ти интервалов). Вычислить плотность в каждом интервале и построить гистограмму. Обозначить модальный интервал и в нем геометрически определить значение моды.
6. Подсчитать по метрической шкале среднее арифметическое значение и среднее взвешенное по распределению. Сравнить их значения.
7. Вычислить дисперсию и среднеквадратическое отклонение третьего признака для групп, выделенных при разных значениях первого признака.
8. Сравнить степень однородности этих групп (п. 7) по значениям коэффициента вариации.
9. Подсчитать энтропию первого признака для двух групп, выбранных по различным значениям второго признака.
10. Вычислить для этих же групп (п. 9) значение коэффициента качественной вариации. Провести сравнительный анализ.

3. АНАЛИЗ ВЗАИМОСВЯЗИ ПРИЗНАКОВ

Условное распределение. Совместное «поведение» двух признаков. Таблица сопряженности. Показатели таблицы сопряженности. Маргинальные частоты. Сравне-

ние структуры условных распределений. Типы задач, решаемых посредством таблиц сопряженности. Типологический синдром. Типологическая группа. Зависимый - независимый признаки. Направленная - ненаправленная связь. Статистическая зависимость - статистическая независимость. Сильная - слабая связь. Меры связи. Функциональная - корреляционная связь. Линейная - нелинейная связь. Локальные - глобальные меры связи. Непосредственная - опосредованная связь. Истинное - ложное значения меры связи.

Независимо от выбранной стратегии анализа (восходящей или нисходящей) и после изучения, условно говоря, «поведения» отдельно взятых признаков, естественным образом возникает необходимость анализа взаимосвязи, взаимодействия между признаками. Будем рассматривать только случай двух признаков. Анализ «поведения» двух признаков — совместного или относительно друг друга — социологу необходим для поиска ответа на вопросы типа: существует ли связь между этими признаками; влияет ли один признак на другой; можно ли, зная значение одного из них, сделать вывод относительно значения другого и т. д. Если гипотезы о взаимосвязях были предварительно сформулированы, то речь может пойти по проверке этих гипотез.

Является очевидным, что поиск ответов на подобные вопросы может осуществляться с помощью ***условных распределений.*** В самом простом случае сравниваются одномерные распределения одного из признаков, полученные для разных совокупностей объектов, на которых второй из признаков принимает одно из своих значений. Возможно также изучать и как бы совместное «поведение» этих признаков.

В качестве исходных для анализа признаков рассмотрим признаки «будущая профессия студента» и «степень удовлетворенности студента учебой». Одномерные распределения этих признаков нам уже известны. Мы будем иметь представление о совместном «поведении» или поведении этих признаков относительно друг друга, если получим так называемую ***таблицу сопряженности*** (корреляционную таблицу). Таковой является таблица 3.3.1. Строки в ней соответствуют шести будущим профессиям (политологи, социологи, культурологи, филологи, психологи и историки), пронумерованным по порядку (они соответствуют профессиональным группам 1, 2, 3, 4, 7, 8 из таблицы 3.2.1), а столбцы - пяти степеням удовлетворенности учебой. Пересечения столбцов и строк образуют ***ячейки*** (клетки) таблицы. В нашем случае число таких ячеек равно 6x5=30. В ячейках таблицы могут содержаться ***значения*** различных показателей. Это - характеристики группы студентов, отнесенных к ячейке, т. е. студентов с определенной будущей профессией, имеющих определенную степень удовлетворенности учебой.

В последней строке представлено распределение (одномерное, простое) студентов по степени их удовлетворенности учебой (частоты обозначены как $\mathbf{n_{0j}}$), а в последнем столбце - распределение студентов по их будущим профессиям ($\mathbf{n_{i0}}$). Для этих частот в контексте анализа таблиц сопряженности есть особое название. Эти частоты называют ***маргинальными частотами,*** и для их обозначения используется, как видите, двойной индекс. В последней строке - маргинальные частоты по столбцам, а в последнем столбце - маргинальные частоты по строкам. Естественно, они совпадают с данными таблиц 3.2.1 и 3.2.2. Сумма маргинальных частот обозначена ($\mathbf{n_{00}}$) и равна 1000, т. е. равна числу наших студентов-гуманитариев.

Любая ячейка таблицы, соответствующая группе объектов, удовлетворяющих условию строки и столбца, может содержать четыре показателя, характеризующих эту группу. К примеру, ячейка (1,2) соответствует 20-ти политологам со второй степенью удовлетворенности учебой (скорее неудовлетворен, чем удовлетворен). Точнее, тем, кто ответил на оба заданных вопроса. Как мы уже знаем, число ответивших может не совпадать с числом опрошенных. Чтобы не было путаницы, будем считать,

что таблица сопряженности получена для некоторой идеальной подвыборки (в нашем случае каждый студент ответил на каждый вопрос). Для обозначения ее объема будем пользоваться понятием - общее число объектов.

Таблица 3.3.1

Таблица сопряженности: абсолютная частота

Будущая профессия студента	**Степень удовлетворенности учебой**					
	1	**2**	**3**	**4**	**5**	$\mathbf{n_{i0}}$
1. Политолог	n_{11}=14	n_{12}=20	n_{13}=31	n_{14}=30	n_{15}=5	n_{10}=100
2. Социолог	n_{21}=30	n_{22}=40	n_{23}=60	n_{24}=60	n_{25}=10	n_{20}=200
3. Культуролог	n_{31}=90	n_{32}=90	n_{33}=60	n_{34}=45	n_{35}=15	n_{30}=300
4. Филолог	n_{41}=31	n_{42}=30	n_{43}=19	n_{44}=15	n_{45}=5	n_{40}=100
5. Психолог	n_{51}=8	n_{52}=10	n_{53}=15	n_{54}=15	n_{55}=2	n_{50}=50
6. Историк	n_{61}=27	n_{62}=110	n_{63}=15	n_{64}=85	n_{65}=13	n_{60}=250
$\mathbf{n_{0j}}$	n_{01}=200	n_{02}=300	n_{03}=200	n_{04}=250	n_{05}=50	n_{00}=1000

Таблица 3.3.2

Таблица сопряженности: относительные частоты

Будущая профессия студента	**Степень удовлетворенности учебой**					
	1	**2**	**3**	**4**	**5**	**Сумма**
1. Политолог	0,14 / 0,07	0,20 / 0,07	0,31 / 0,15	0,30 / 0,12	0,05 / 0,10	1,00
2. Социолог	0,15 / 0,15	0,20 / 0,13	0,30 / 0,30	0,30 / 0,24	0,05 / 0,20	1,00
3. Культуролог	0,30 / 0,45	0,30 / 0,30	0,20 / 0,30	0,15 / 0,18	0,05 / 0,30	1,00
4. Филолог	0,31 / 0,16	0,30 / 0,10	0,19 / 0,09	0,15 / 0,06	0,05 / 0,10	1,00
5. Психолог	0,16 / 0,04	0,20 / 0,03	0,30 / 0,08	0,30 / 0,06	0,04 / 0,04	1,00
6. Историк	0,11 / 0,13	0,44 / 0,37	0,06 / 0,08	0,34 / 0,34	0,05 / 0,26	1,00
Сумма	/ 1,00	/ 1,00	/ 1,00	/ 1,00	/ 1,00	

Для политологов, имеющих вторую степень удовлетворенности учебой, абсолютная частота равна $\mathbf{n_{12}}$. Кроме нее в ячейку (1,2) можно поместить и значения других показателей, а именно относительных частот либо в долях (частости), либо в процентах. При этом таких частот может быть три. Назовем абсолютную частоту первым показателем в ячейке таблицы сопряженности, и будем исходить из того, что относительные частоты рассчитываются в долях. Тогда второй показатель будет равен доле этих $\mathbf{n_{12}}$ студентов в общем числе $\mathbf{n_{00}}$, студентов-гуманитариев. Третий показатель - доля этих же $\mathbf{n_{12}}$ студентов среди $\mathbf{n_{10}}$ студентов-политологов. Четвертый - доля этих же $\mathbf{n_{12}}$ студентов среди $\mathbf{n_{02}}$ студентов, степень удовлетворенности учебой которых равна двум.

Теперь запишем все это в общем виде (в виде формул) для объектов любой природы и для любой (i, j)-й ячейки таблицы сопряженности. Число объектов, удовлетворяющих условию i-й строки и j-го столбца, равно $\mathbf{n_{ij}}$, общее число объектов равно $\mathbf{n_{00}}$. Маргинальные частоты по столбцам - $\mathbf{n_{0j}}$, а маргинальные частоты по строкам - $\mathbf{n_{i0}}$. Символ «нуль» обозначает, что по тому индексу, на месте которого он стоит, проведено как бы суммирование или усреднение или расчеты проведены без учета некоторого признака. Это очень удобный способ для обозначений частот разного вида, возникающих при анализе таблицы сопряженности. Вместо этого символа можно использовать и другой, например, точку или звездочку. «Точка», «звездочка», «нуль» - общепринятые в литературе символы для обозначения маргинальных частот.

Таким образом, (i, j)-й ячейке таблицы сопряженности можно поставить в соответствие четыре показателя:

1. $\mathbf{n_{ij}}$ - число объектов, удовлетворяющих условию i-й строки и j-го столбца;
2. $\mathbf{n_{ij}/\ n_{00}}$ - доля их в общей совокупности объектов;
3. $\mathbf{n_{ij}/\ n_{i0}}$ - доля их в совокупности объектов, удовлетворяющих условию строки;
4. $\mathbf{n_{ij}/\ n_{0j}}$ - доля этих же объектов в совокупности объектов, удовлетворяющих условию столбца.

Социолог анализирует «поведение» одного признака относительно другого с помощью двух последних показателей. В таблице 3.3.2 приведены в каждой ячейке значения этих двух показателей для нашей задачи. Над чертой в ячейке доля по строке, а под чертой - доля по столбцу. На основе этих данных социолог может решать два типа задач.

Во-первых, он может сравнивать ***структуру*** «удовлетворенности учебой» в различных профессиональных группах студентов. Мы упомянули новый в нашем курсе термин «структура». В самом простом случае под структурой «чего-то» понимается совокупность элементов этого «чего-то» и взаимосвязи между этими элементами.

Это вам знакомо. В нашем случае элементами являются различные степени удовлетворенности учебой, а в качестве взаимосвязи между ними выступает различие в «долях», соответствующих этим степеням. Для того чтобы представить эти структуры графически, построим на одном и том же графике эмпирические кривые распределения по удовлетворенности учебой отдельно для каждой профессиональной группы студентов-гуманитариев.

На рис. 3.3.1 изображены шесть эмпирических кривых распределения, соответствующих шести профессиональным группам. На горизонтальной оси отложены на равном расстоянии пять степеней удовлетворенности. Чтобы построить кривую распределения для политологов (первая наша профессиональная группа), по вертикальной оси откладываем следующие значения (0,14, 0,20, 0,31, 0,30, 0,05) из первой строки таблицы 3.3.2. Это доли политологов с соответствующей степенью удовлетворенности (от 1 до 5) среди всех политологов. Аналогично поступаем и в случае остальных профессиональных групп. К примеру, чтобы построить кривую распределения для студентов-психологов, по вертикали откладываем следующие значения (0,16, 0,20, 0,30, 0,30, 0,04) соответственно пяти степеням удовлетворенности учебой.

Чисто визуально из рис. 3.3.1 можем сделать следующие выводы. Структура удовлетворенности «похожа» у политологов, социологов и психологов. Эти группы образуют как бы один ***типологический синдром,*** составляют одну и ту же ***типологическую группу*** по структуре удовлетворенности. Структура удовлетворенности примерно одинакова у культурологов и филологов. Это уже второй типологический синдром. Таким образом, можно утверждать наблюдаем наличие трех типологических синдромов при анализе структуры удовлетворенности. Третий из них - специфическая и отличная от других структура удовлетворенности учебой студентов-

историков. Эти синдромы, типологические образования и есть специфические эмпирические закономерности, требующие от социолога объяснения. В целом можно констатировать, что будущая профессия студента влияет на удовлетворенность учебой или детерминирует эту удовлетворенность. На вопрос, каким образом, мы тоже ответили пока без каких-либо количественных оценок. Как видите, в этом случае визуализация распределений имеет для социолога огромное значение.

Выше упоминали два типа задач, решаемых с помощью таблицы сопряженности. Первый тип мы с вами рассмотрели. Формально мы анализировали третий показатель таблицы сопряженности. Другой из этих типов задач для нашего примера заключается в сравнении профессиональной структуры в различных по степени удовлетворенности учебой группах студентов. На рис. 3.3.2 изображены пять эмпирических кривых распределения в соответствии с этими группами. Для построения этих кривых используем четвертый показатель таблицы сопряженности. В таблице 3.3.2 значения этого показателя находятся под чертой. Для того чтобы построить, к примеру, эмпирическую кривую распределения студентов по их будущим профессиям для третьей группы по степени удовлетворенности (частично удовлетворенные и частично неудовлетворенные), из таблицы 3.3.2 выделим столбец со значениями (0,16, 0,30, 0,30, 0,10, 0,08, 0,08). Это доли шести профессиональных групп в совокупности удовлетворенных учебой на тройку. Аналогичным образом строятся и другие четыре кривые распределения.

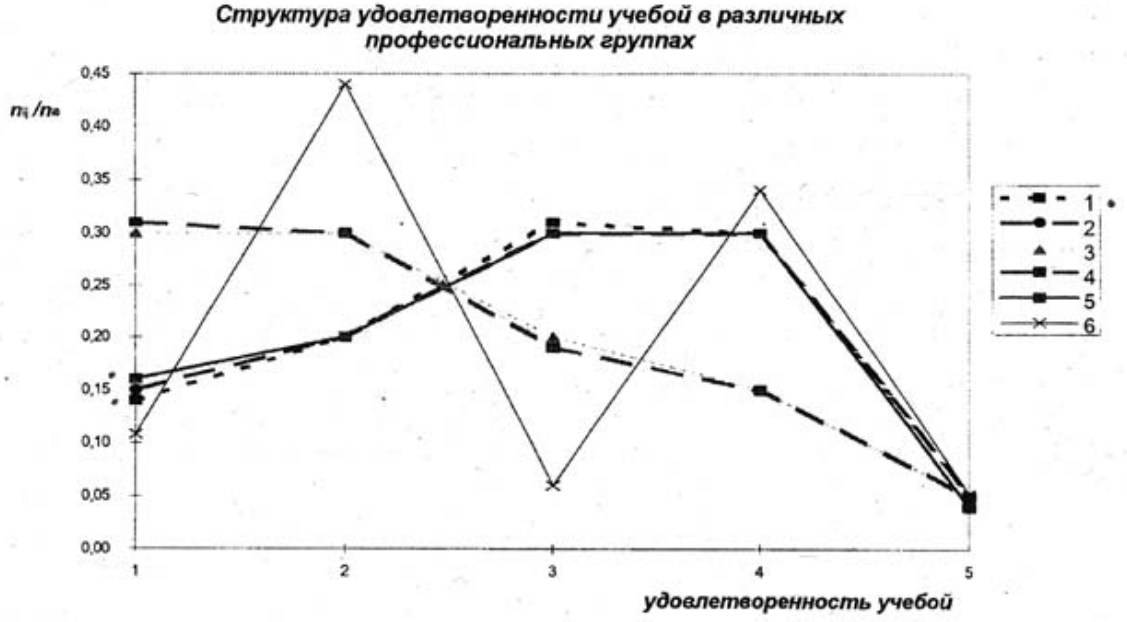


Рис. 3.3.1.

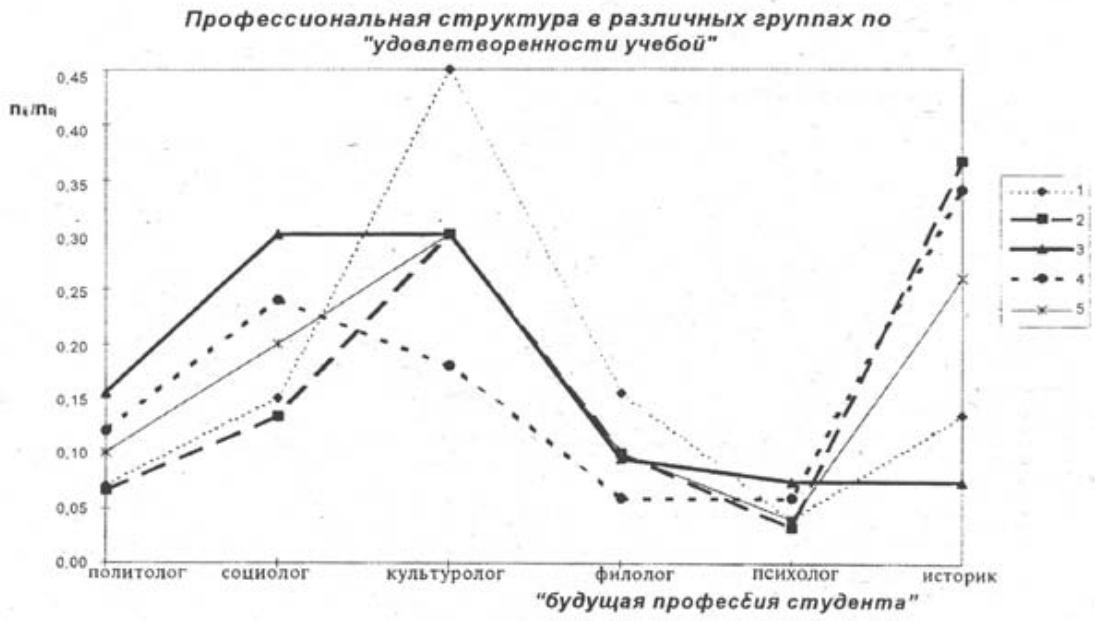


Рис. 3.3.2.

Из визуального сравнения пяти построенных нами эмпирических кривых распределения видим следующее. Похожесть профессиональных структур наблюдается только для третьей и четвертой групп по удовлетворенности учебой. Практически в каждой группе, кроме этих двух, по удовлетворенности своя собственная профессиональная структура. Из этого делаем следующий вывод, что признаки «будущая профессия» и «удовлетворенность учебой» статистически (по данным) связаны. Обратите внимание, что формально можно говорить о влиянии удовлетворенности на профессию, но содержательно это не имеет никакого смысла.

Это пример того, как выбор «языка» интерпретации эмпирической закономерности обусловлен содержанием признаков. В первом типе задач «язык» влияния, «язык» детерминации имеет смысл, а во втором типе не имеет смысла. Соответственно в первом случае имеет смысл понятие направленной связи. Поэтому иногда очень важно заранее определить, какой из признаков может содержательно зависеть от другого. Отсюда возникают понятия ***зависимый*** (целевой) и ***независимый*** признак. Дихотомия ***«направленная - ненаправленная»*** связь является важной в понимании связи.

Деление на зависимые - независимые признаки в социологии не всегда содержательно обосновано. Зачастую такое деление необходимо в процессе анализа и носит функциональный характер. В том смысле, что один и тот же признак независимо от его содержания в одной задаче может выступать в роли зависимого, а в другой - в роли независимого. Причем в рамках одного и того же исследования. Разумеется, присутствующая в каждом опросе «объективка» (пол, возраст, образование, происхождение и т. д.) порождает признаки, трактуемые как независимые.

Если вернуться к рис. 3.3.1 и к рис. 3.3.2, то можно заметить следующее. Представим себе, что все кривые на каждом из рисунков похожи между собой. Что это означает для социолога? Во-первых, это значит, что профессиональная структура в группах студентов с различной степенью удовлетворённости учебой одинакова и не зависит от этой степени. При этом она (структура) такая же, как и профессиональная структура для всей совокупности студентов-гуманитариев (маргинальные частоты по строкам). Во-вторых, это значит, что структура удовлетворенности во всех профессиональных группах одинакова и не зависит от будущей профессии студента. При этом эта структура такая же, как во всей совокупности (маргинальные частоты по

столбцам). Тогда связь между феноменами «профессия» и «удовлетворенность» отсутствует, ***статистическая связь*** не наблюдается. Наши признаки ***статистически независимы.***

Нетрудно догадаться, что в исследованиях такая ситуация практически не встречается, и не потому, что отсутствие связи не наблюдается, а совсем по другим причинам. Основная причина - специфика наших социологических данных. Это их неустойчивый характер. Например, это проявляется в неточности измерения того же феномена, как удовлетворенность учебой. Причин тому множество. Это и несовершенство методик измерения, и неустойчивость ответов респондента, и плохая выборка. Ясно одно, всегда имеет место влияние многих случайных и неслучайных факторов на конкретные значения изучаемого нами признака. С неслучайными факторами социолог может бороться, а случайные будут иметь место всегда. Поэтому социолог делает выводы с учетом этой ситуации. Задается уровнем «ошибиться». Статистическая независимость констатируется не в идеальном случае, а в случае, близком к идеальному.

Представим себе противоположную ситуацию, когда на каждом из рисунков все кривые непохожи, несхожи. Для социолога это означает, что в каждой группе с разной степенью удовлетворенности учебой своя собственная профессиональная структура. В каждой профессиональной группе своя собственная структура удовлетворенности. Из этого следует, что будущая профессия студента связана с его удовлетворенностью учебой, наблюдается сильная ***статистическая зависимость.*** Естественно, такая ситуация в исследованиях тоже практически не встречается.

Реальные рисунки трудно поддаются визуальной интерпретации. К тому же в исследовании их бывает очень много. Отсюда и возникает необходимость в количественных оценках степени взаимосвязи между признаками, в определении, ***сильное*** или ***слабое*** влияние признаков друг на друга. Это можно сделать с помощью различных мер взаимосвязи. Мы подошли к важным понятиям ***меры связи,*** или ***коэффициенты связи.*** Таких мер много, так как много различных интерпретаций понятия «связь». Другими словами, связь может пониматься по-разному. Это во-первых. Во-вторых, даже в рамках одного и того же понимания связи существуют различные способы ее математической формализации. Отдельно взятый коэффициент - математическая формализация некоторого понимания связи.

То, что нужны некоторые количественные оценки степени похожести эмпирических кривых распределения, не вызывает теперь у вас никакого сомнения. Но это только один контекст, одна из интерпретаций понимания связи. Прежде чем рассмотреть различные коэффициенты связи, введем дихотомические пары понятий, без которых невозможно перейти к эмпирической интерпретации понятия «связь». Каждая интерпретация или контекст порождает свою собственную группу коэффициентов связи. Эти дихотомические пары для социолога составляют понятийный аппарат при использовании в анализе понятия «связь». Некоторые из этих пар были упомянуты выше: *зависимый признак - независимый, направленная связь - ненаправленная, статистическая зависимость - независимость, сильная (тесная) связь - слабая.*

Коротко поясним содержательный смысл еще нескольких пар понятий. При этом будем упоминать коэффициенты связи (пока их названия, принятые в литературе), которые будут введены в следующем разделе. Итак, следующая пара понятий: ***функциональная связь - корреляционная связь.*** Из школьной математики вы прекрасно знаете, что функциональной связью между двумя признаками называется такая связь, когда одному и тому же значению одного признака соответствует одно или несколько значений другого. Геометрически - это красивые плавные кривые (прямая, парабола, синусоида и т. д.) или кривые с точкой разрыва (гипербола). Функциональные связи в социологии встречаются в основном при работе с данными первого типа. Примером функции является и любой аналитический индекс. При рассмотрении свя-

зи между двумя признаками в рамках других типов информации наблюдается другая картина - одному и тому же значению признака соответствует целое распределение значений по другому из признаков. Такая связь называется корреляционной (точнее, стохастической, но мы такие тонкости, как различие стохастических и корреляционных связей, рассматривать не будем). Эти связи между двумя признаками геометрически могут быть изображены в виде облаков точек в двумерном пространстве, т. е. на плоскости.

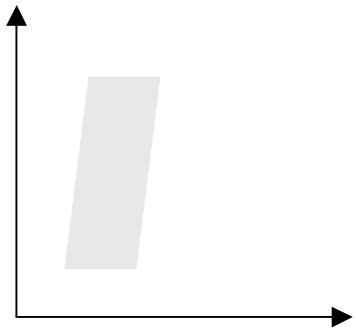


Рис. 3.3.3 Сильная связь

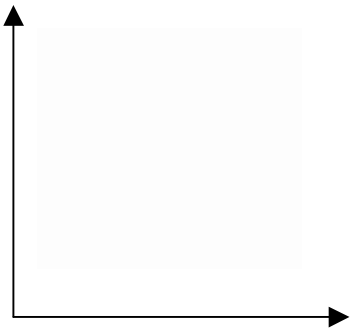


Рис. 3.3.4 Слабая связь

Корреляционная связь может быть сильной (рис. 3.3.3) и слабой (рис. 3.3.4). В первом случае облако точек имеет четкую конфигурацию, четкую закономерность. Если признаки имеют метрический уровень измерения, то можно сказать, что с ростом значений одного признака растет в среднем и значение другого. Здесь наблюдаем ***линейную связь.*** Эта закономерность может быть описана посредством прямой линии, которая называется ***линией регрессии.*** Разумеется, корреляционная связь может быть и ***нелинейной,*** т. е. описываться не прямыми.

Для нас важно, что корреляционные связи могут быть описаны с помощью функциональных. Другими словами, социологу правомерно ставить вопрос, насколько корреляционная связь отличается от заданной им (в виде гипотезы) функциональной. С аналогичной ситуацией мы уже сталкивались. Практически все коэффициенты качественной вариации основаны на оценке степени отклонения от равномерного распределения (от прямой линии).

Социолог сталкивается с необходимостью задавать или выбирать функциональные зависимости при работе с любым из пяти типов информации. При работе с динамическими рядами главная задача - построить, подобрать функцию, описывающую этот ряд. Многие математические методы предполагают задание характера зависимости изучаемых признаков. Правда, из этого не следует, что мы всегда найдем функцию, подходящую для описания эмпирической закономерности.

Существует мера связи в предположении, что корреляционная связь носит линейный характер, и признаки имеют метрический уровень измерения. Такая мера называется коэффициентом линейной связи Пирсона.

Целесообразно также использование такой пары понятий, как ***глобальные - локальные*** меры связи. Эта пара понятий необходима для условного обозначения следующей ситуации. Вернемся к таблице сопряженности для нашего случая. Как было отмечено, определить связь между будущей профессией студента и удовлетворенностью учебой можно, сравнивая их условные распределения. В этом случае речь идет как бы о связи этих двух признаков в целом. Меры, отражающие эту целостность, можно определить условно как меры «глобального» характера для таблицы сопряженности. К такого рода мерам относятся коэффициенты, основанные на величине «хи-квадрат» и Гудмена-Краскала.

В то же время можно поставить вопрос о связи следующим образом. Например, связана ли самая низкая удовлетворенность учебой со второй профессией (социолог).

Тогда речь идет условно как бы о связях в локальном смысле. Для таких случаев существуют также коэффициенты связи. Это такие коэффициенты, как коэффициент Юла, показатели детерминации.

Вместо рассмотренной пары направленная связь - ненаправленная можно пользоваться терминами: ***симметричная связь - асимметричная.*** При вычислении направленных коэффициентов связи между признаками X и Y, как правило, оказывается, что значение коэффициента для $X \rightarrow Y$ не равно значению для $X \leftarrow Y$. Два признака неравноправны, их нельзя формально поменять местами. Отсюда возникают асимметричные коэффициенты. Они не всегда удобны для использования в сложных математических методах. Потому при двух асимметричных коэффициентах всегда существует третий, как бы их усредняющий. Мы столкнемся с тройкой мер Гуттмана и с тройкой мер Гудмена - Краскала.

Перейдем к рассмотрению взаимосвязанных пар понятий, таких, как ***непосредственная связь - опосредованная, истинное*** (значение коэффициента) - ***ложное.*** Первая пара понятий важна при интерпретации количественного значения коэффициента связи. Здесь необходимо отметить, что по таким значениям не всегда можно говорить о силе связи (сильная - слабая). В ряде случаев просто констатируется наличие или отсутствие определенным образом понимаемой связи. Если по конкретному значению коэффициента мы видим, что связь есть, то это вовсе не означает существования в реальности непосредственной связи между двумя изучаемыми признаками, а может означать наличие опосредованной связи. Отсюда вторая пара понятий: истинное значение - ложное. В литературе тому есть множество примеров. Например, в США за 1870—1910 годы было установлено наличие связи между заработной платой учителей и потреблением вина. Это пример ложной связи. Ибо она была опосредована тем, что в эти годы наблюдался промышленный бум и рост заработной платы и тем самым рост потребления вина во всех группах населения. В нашем случае можно сказать, что связь между будущей профессией студента и удовлетворенностью учебой есть. Но она может носить ложный характер, т. е. опосредована другими признаками. Например, социальным происхождением, успеваемостью, удовлетворенностью жизнью, уверенностью в завтрашнем дне и т. д.

Возможна и другая ситуация, когда значение коэффициента связи указывает на ее отсутствие, а на самом деле связь существует. Пример приведем в следующем разделе книги для случая таких признаков, как удовлетворенность собой и удовлетворенность жизнью.

Еще несколько слов о статистической зависимости - статистической независимости. Это очень важные понятия. Вернемся опять к нашей таблице сопряженности и задаче сравнения условных распределений. Выше, исходя из элементарного здравого смысла, мы пришли к необходимости использования направленных мер связи для определения различия в структурах распределения. Тем самым для определения: наблюдается ли статистическая зависимость между будущей профессией студента и удовлетворенностью учебой. Но для определения статистической зависимости можно исходить и из другой модели, из других соображений. Поставим вопрос так. Какая величина может стоять в ячейке таблицы сопряженности, если эти признаки статистически независимы? Разумеется, такой вопрос правомерен. При этом маргинальные частоты (одномерные, простые) нам известны по нашей выборке.

Рассмотрим, к примеру, ячейку (2,1). Она соответствует будущим социологам, неудовлетворенным учебой. Статистическую независимость признаков «будущая профессия» и «удовлетворенность учебой» можем понимать следующим образом. Доля неудовлетворенных учебой социологов среди всех студентов-социологов равна доле не удовлетворенных учебой студентов среди всех студентов-гуманитариев. Ведь такое понимание связи не должно вызывать у вас неприятия, ибо не противоречит здравому смыслу социолога. Тогда в ситуации статистической независимости легко

определяется то значение, которое должно стоять в нашей ячейке. Оно вычисляется исходя из упомянутой выше пропорции. К ней мы вернемся при рассмотрении мер связи, основанных на так называемой величине «хи-квадрат».

Многие коэффициенты связи как раз и определяют отклонение реальных частот (того, что получено по выборке) от частот как бы теоретических, т. е. вычисленных по той же таблице, но для случая статистической независимости.

И, наконец, обратим внимание еще на одну пару понятий. Социолога интересует связь между признаками для выявления причинно-следственных отношений между признаками. Поэтому он изучает связи всегда в контексте: влияет - не влияет; детерминирует - не детерминирует; увеличивает информацию - не увеличивает; улучшает прогноз - не улучшает и т. д. После всех наших предыдущих рассуждений является очевидным, что наличие корреляционной связи не говорит о причинности [3. с. 72—119; 11. с. 43—63]. И в то же время для причинного анализа невозможно обойтись без изучения корреляционных связей. Термином «причинный анализ» принято обозначать специфический класс математических методов. Вместе с тем проблема причинности в нашей науке очень интересная, сложная область, которую нельзя свести только к классу математических методов.

Итак, мы познакомились с дихотомическими парами понятий, которые важны для изучения и понимания связи, т. е. для эмпирической интерпретации понятия «связь». Они таковы:

причинная - корреляционная; функциональная - корреляционная; направленная - ненаправленная; локальная - глобальная; истинная - ложная; статистическая зависимость - статистическая независимость; симметричная - асимметричная; непосредственная - опосредованная; линейная - нелинейная.

Коэффициенты связи, меры связи бывают не только парные (мы будем рассматривать только такие), но и частные, множественные. Различают коэффициенты для номинального, порядкового, метрического уровня измерения. Сами таблицы сопряженности бывают разные. Они бывают и многомерные, если сопрягаются несколько признаков, и тогда их называют таблицами с несколькими входами. Очень интересной в социологии является таблица сопряженности квадратного вида (число строк равно числу столбцов), когда сопрягается признак с самим собой. Она возникает в ситуации панельного исследования. Представим себе, что тех же студентов-гуманитариев мы опросили повторно через пару лет. Тогда таблица для двух признаков, например, «уверенность в завтрашнем дне в 1997 году» и «уверенность в завтрашнем дне в 1999 году», позволит изучить степень изменчивости такой уверенности. Для анализа таких таблиц сопряженности существуют специфические меры связи.

Задание на семинар или для самостоятельного выполнения

1. На основе той же самой матрицы данных составить таблицу сопряженности между первым (номинальная шкала) и вторым (порядковая шкала) признаками. В каждой ячейке таблицы подсчитать значение четырех показателей: абсолютную частоту и относительные частоты в долях (частости) по всем объектам, по строке и по столбцу.

2. На одном и том же рисунке построить эмпирические кривые распределения по первому признаку для различных групп объектов, выделенных по отдельным значениям второго признака. Сравнить эти кривые и сделать выводы о характере связи двух признаков, о наличии типологических синдромов.

3. На одном и том же рисунке построить эмпирические кривые распределения по второму признаку для различных групп объектов, выделенных по отдельным значениям первого признака. Сравнить эти кривые и сделать выводы о характере связи

двух признаков, о наблюдаемых эмпирических закономерностях.

4. МЕРЫ СВЯЗИ, ОСНОВАННЫЕ НА ПОНЯТИЯХ «СТАТИСТИЧЕСКАЯ ЗАВИСИМОСТЬ» И «ДЕТЕРМИНАЦИЯ»

Две логические схемы использования коэффициентов связи. Локальные меры связи для таблиц сопряженности. Коэффициент Юла. Понятие детерминации. Интенсивность и емкость детерминации. Оценки вероятности. Истинное - ложное значение мер связи. Понятие о величине χ^2 (хи-квадрат). Коэффициент взаимной сопряженности Е. Пирсона. Значимость значений коэффициентов связи. Доверительный интервал.

Рассмотренные дихотомические пары понятий, составляющие контекст для эмпирической интерпретации понимания связи, естественным образом привели нас к выводу о необходимости существования большого количества коэффициентов парной связи. Каждая мера связи (каждый коэффициент связи) вводится таким образом, чтобы его значения изменялись либо от нуля до единицы, либо от минус единицы до единицы. Это единственное, что объединяет все, коэффициенты. Перед социологом всегда стоит трудный вопрос, связанный с тем, как понимать связь и какой коэффициент выбрать для изучения взаимосвязи между признаками. Иногда возникает иллюзия, что, получив значения всевозможных коэффициентов и сравнив эти значения между собой, можно сделать достоверный вывод о силе связи между признаками. Дело в том, что сравнивать имеет смысл только коэффициенты, основанные на одном и том же понимании связи.

Обычно раздражение социолога-пользователя вызывает и то, что нельзя сравнивать силу связи в разных исследованиях по значениям коэффициентов. Если в одном исследовании коэффициент равен 0,5, а в другом тот же коэффициент для тех же признаков 0,6, нельзя утверждать, что второе больше первого. Ведь социолог, анализируя связь, всегда ищет ответы на вопросы: «Насколько влияет/не влияет...?», «Насколько зависит/не зависит...?». Коэффициенты же зачастую на эти вопросы не отвечают. У них свой язык понимания связи, который необходимо понять. Только тогда появляется возможность использования их для ответа на подобные вопросы.

Для того чтобы правильно пользоваться каким-нибудь коэффициентом, необходимо прежде всего знать все его возможности и не требовать от него того, чего он не может дать социологу. В социологических исследованиях сами значения коэффициентов, как правило, бывают маленькими. Наблюдается такая странная картина, когда все анализируемые признаки друг с другом связаны, но очень слабо (по значениям мер взаимосвязи). Почему это происходит - понятно. Мы с помощью парных связей рассматриваем непосредственные связи между двумя признаками, а в социологии все опосредовано. Другими словами, на нашу пару признаков влияют множество других. Что это за признаки, не всегда известно. Поэтому использование отдельно взятого коэффициента эффективно только в сравнительном контексте и только в рамках одного исследования. Например, возможны две логические схемы использования парных коэффициентов связи.

Первая состоит в следующем. Из всей совокупности признаков, связи между которыми интересуют социолога, выделяется какой-то важный, главный, ***зависимый, целевой*** признак, и рассматриваются его парные связи с остальными. В самом простом случае последние считаются как бы независимыми друг от друга и влияющими в разной степени на целевой. Вычисляются значения коэффициента и по этим значениям проводится процедура ранжирования всех независимых признаков по степени их влияния на целевой. Затем на основе сугубо качественного анализа отбираются из независимых наиболее тесно связанные с целевым. Этот прием чисто практический и теоретически может быть и необоснован. К сожалению, социологу на каждом шагу приходится идти на подобные нарушения. Такая логическая схема анализа может

вывести социолога к необходимости формирования новых гипотез о ***причинно-следственных отношениях*** между признаками.

Вторая схема возникает в ситуации невозможности (содержательной бессмысленности) выделения целевого из всей совокупности анализируемых признаков. Тогда вычисляются значения коэффициента связи для всевозможных пар признаков. С помощью задания некоторого ***порога*** (значения коэффициента) отсекаются все связи со значением коэффициента, который меньше этого порога. Строится граф структуры взаимосвязей, где вершины - признаки, а ребра - связь между ними. Пусть у нас с вами каких-то шесть признаков и вычислены значения какого-то коэффициента. На рис. 3.4.1 и на рис. 3.4.2 приведены два графа.

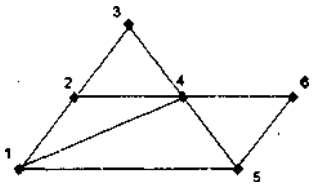


Рис. 3.4.1 Граф связи

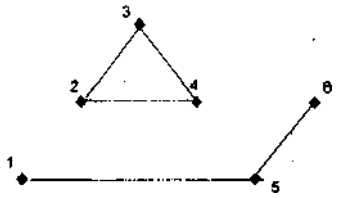


Рис. 3.4.2 Граф связи

Первый из них получился с большим количеством связей, потому что задали маленькое значение порога. Второй граф получился с очень маленьким числом связей, потому что задали большое значение порога. Значения коэффициентов не имеет особого смысла приводить. Нам важен только содержательный смысл этой процедуры. На первом графе могут быть изображены и несущественные связи, а во втором наоборот - существенные могли быть потеряны. Независимо от содержания признаков, принцип выбора порога всегда носит итеративный характер и критерии всегда качественные. Такая логическая схема может вывести социолога к формированию гипотез о социальных факторах. Ибо на втором графе наблюдаем, к примеру, два ***факторных синдрома,*** т. е. две группы взаимосвязанных признаков, что является основой для формирования ***индексов.***

Эти логические схемы порождены двумя самыми простыми задачами изучения структуры взаимосвязи совокупности признаков. Они опираются на парные коэффициенты связи, к рассмотрению которых мы и переходим. При этом перед нами стоит трудная задача. С одной стороны, даже в социологической литературе существует множество работ с описанием коэффициентов связи [3, 8, 9, 11]. С другой стороны, студенты-социологи с большим трудом воспринимают такого рода материал. С учетом этой ситуации мы будем рассматривать только некоторые коэффициенты. Основное внимание обратим только на то, на каком ***понимании связи*** основана та или иная группа коэффициентов, и на специфику ***языка анализа*** парных взаимосвязей между признаками. Математических обоснований касаться не будем, оставляя их для освоения на последующих этапах вашего образования.

Локальные меры связи

Речь идет об анализе данных, представленных в виде так называемых таблиц сопряженности вида (2×2). Предположим, что необходимо проанализировать связь между первой профессией (будущая профессия студента - политолог) и четвертой степенью удовлетворенности учёбой (скорее удовлетворенные, чем нет). В этом случае удобно говорить на языке изучения связи двух свойств. В нашем случае первое свойство - быть политологом, второе - быть удовлетворенным учебой на четыре балла. Отдельно взятый студент (в других задачах это любой другой эмпирический объект) либо обладает одним из этих двух свойств, либо обладает одновременно двумя свойствами, либо не обладает никаким из этих свойств.

Из нашей предыдущей таблицы 3.3.1 видим, что будущих политологов, удовле-

творенность учебой которых равна четырем баллам, было 30 человек. Студентов, обладающих первым свойством, всего 100 человек, а обладающих вторым свойством - 250. Таблица 3.4.1 представляет собой таблицу вида (2×2) для наших двух свойств. В ячейках таблицы в скобках приведены условные обозначения абсолютных частот (a, b, c, d). В данном случае можно обойтись без индексов. Маргинальные частоты обозначены как суммы этих четырех частот.

Таблица 3.4.1

Таблица сопряженности для первой профессии и четвертой степени удовлетворенности

	Удовлетворенные учебой на «четыре»	**«Остальные»**	**Итого**
Будущие политологи	30 (a)	70 (d)	100 (a+d)
«не политологи»	220 (c)	680 (b)	900 (c+d)
Итого	250 (a+c)	750 (d+b)	1000 (a+b+c+d)

Одним из языков анализа связи между этими свойствами является поиск ответа на вопрос: наблюдается ли статистическая зависимость между этими свойствами. Если наблюдается статистическая независимость У (удовлетворенные учебой на «четыре») от П (политологи), то 30/250 (доля удовлетворенных учебой политологов среди всех удовлетворенных учебой на четыре балла) должно равняться 70/750 (доля «остальных» политологов среди всех «остальных»). То же самое запишем в общем виде:

$$\frac{a}{a+c} = \frac{d}{d+b}$$

Из этого следует, что $a(d + b) = (a + c)d \rightarrow db = cd$. Тогда разность ab —cd можно использовать как меру отклонения от статистической независимости. Такое же соотношение получим, если будем рассуждать по-другому. Если статистическая независимость П от У наблюдается, то, 30/100 (доля удовлетворенных политологов среди политологов) должно равняться 220/900 (доля удовлетворенных «не политологов» среди всех «не политологов»).

На этой разности и основан коэффициент Юла (G. Yule), который имеет следующий вид:

$$Q = \frac{ab - cd}{ab + cd}$$

Знаменатель введен для того, чтобы значения этого коэффициента изменялись от -1 до +1. Если вы видите коэффициенты двухэтажные (со знаменателями), то очень часто (но не всегда) наличие знаменателя служит как бы для нормирования интервала изменения значений коэффициента. Содержательный смысл меры связи, как правило, передает числитель. Рассмотрим свойства (поведение) этого коэффициента:

1. Он равен единице либо когда c=0 /схема 3.4.1 а)/, либо d = 0 /схема 3.4.1 б)/. В первом случае все «не политологи» относятся к «остальным» по удовлетворенности. Обратное утверждение неверно. Во втором случае все политологи удовлетворены учебой на 4 балла. Опять же обратное утверждение будет неверным.

2. Он равен минус единице, если a=0 /схема 3.4.1 в)/ или b = 0 /схема 3.4.1 г)/. В первом случае все политологи относятся к «остальным» по удовлетворенности. Во втором случае все «не политологи» удовлетворены учебой на четыре балла. Обратные утверждения неверны.

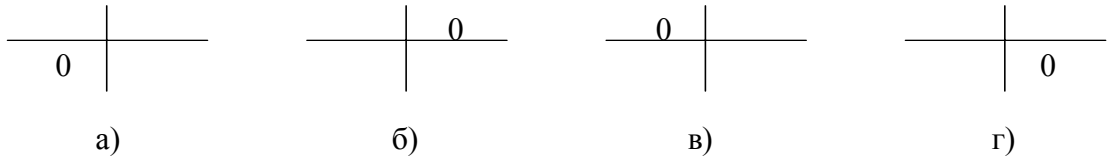


Схема 3.4.1

3. Коэффициент равен нулю, если *ab = cd,* т. е. в случае статистической независимости наших изучаемых свойств.

В нашем случае коэффициент равен 0,14. Естественным образом, возникает вопрос, каким будет значение коэффициента для генеральной совокупности. Ведь пока мы получили только оценку связи по выборочной совокупности. Значение коэффициента небольшое, но отличное от нуля, поэтому возникает другой вопрос. Значимо ли это отличие от нуля или мы получили ненулевое значение случайно? Если это отклонение незначимо, то наблюдается статистическая независимость наших свойств (быть политологом и быть удовлетворенным учебой на четыре балла). И наоборот, если это отклонение значимо, то имеем случай статистической зависимости. Для определения значимости и для определения «истинного» значения (для генеральной совокупности) необходим аппарат математической статистики, а именно аппарат проверки ***статистических гипотез.*** Их не следует путать с содержательными гипотезами исследования. К этому вопросу мы вернемся несколько позже после введения так называемой статистики **хи-квадрат.**

Рассмотрим использование меры Юла в сравнительном контексте. Пусть целевое свойство - «быть удовлетворенным учебой на четыре балла». Попытаемся определить, какая из будущих профессий теснее связана с этим свойством, сильнее влияет на подобную удовлетворенность. По данным, представленным в таблице 3.3.1, сформируем таблицы сопряженности вида (2×2) для подсчета шести значений для шести будущих профессий. Так как для политологов значение коэффициента уже было получено по таблице 3.4.1, то ниже на схеме 3.4.2 приведены таблицы для оставшихся пяти будущих профессий. В этих таблицах приведены только абсолютные частоты. Целевой признак обозначен как (У). (+У) и означает обладать свойством «удовлетворенности учебой на четыре балла», а (-У) - не обладать, т. е. остальные варианты удовлетворенности учебой.

+У	-У
60	140
190	610

(2) социологи

+У	-У
45	225
205	495

(3) культурологи

+У	-У
15	85
235	665

(4) филологи

+У	-У
15	35
235	715

(5) психологи

+У	-У
85	165
165	585

(6) историки

Схема 3.4.2. Таблицы сопряженности «удовлетворенность учебой на 4 балла» с будущими профессиями студентов

Для политологов коэффициент Юла был равен Q_1=0,14. Для социологов Q_2=0,16, так как

$$Q_2 = \frac{60 \times 610 - 190 \times 140}{60 \times 610 - 190 \times 140} = \frac{36600 - 26600}{36600 + 26600} = \frac{10000}{63200} = 0{,}16$$

Аналогичным образом вычисляются значения коэффициента для культурологов, филологов, психологов и историков. Соответственно получим следующие значения:

Q_3=0,40; Q_4 =--0,33; Q_5 =0,13; Q_6=--0,29.

Таким образом, если не учитывать, прямая (значения коэффициента положительные) или обратная (значения коэффициентов отрицательные) связь, наши шесть профессий по степени влияния на удовлетворенность упорядочиваются следующим образом:

$|Q_3|>|Q_4|>|Q_6|>|Q_2|>|Q_1|>|Q_5|$

Свойства «быть культурологом» и «быть филологом», скорее всего, связаны со свойством «удовлетворенность учебой на четыре балла» и влияют на него. Свойства «быть психологом» и «быть политологом», скорее всего, не влияют. От них удовлетворенность учебой не зависит. Еще раз хочется напомнить, в каком смысле «влияет», в каком смысле «зависит». Пока только в смысле статистической зависимости. Почему мы говорим «скорее всего»? Потому что по формальным критериям может оказаться, например, что все значения коэффициентов незначимо отличаются от нуля. Полученный результат ранжирования — лишь контекст для формирования новых содержательных гипотез и усложнения моделей изучения связи.

Понятие детерминации

Для анализа локальной связи можно использовать и язык детерминации [14]. Правило ***«если С, то У»*** называется ***детерминацией.*** Термин «determinatio» был введен в 1900 году в биологии и обозначает ситуацию, когда одно свойство («быть будущим социологом (С)») оказывает влияние на другое («быть удовлетворенным учебой (У)»). Такое влияние обозначается, «С→У». Детерминация имеет две основные характеристики: ***интенсивность*** **I** (С→У) детерминации и ***емкость С*** (С→У) детерминации. Формально — это условные частоты. В наших обозначениях эти характеристики равны:

$$I\ (C\rightarrow У) = \frac{a}{a+d};\ C\ (C\rightarrow У) = \frac{a}{a+c}$$

Если значения этих характеристик выразить в процентах (что очень удобно для интерпретации), то для нашего примера (см. таблицу 3.3.1 или первую табличку на схеме 3.4.2):

$$a = 60;\ a + d=200;\ a + c=250.$$

Тогда I (С→У)=30%, а С (С→У)=24%. Из первого значения делаем вывод, что из числа студентов, обладающих свойством С (быть социологом), 30% обладают свойством У (быть удовлетворенным). Интенсивность выражает как бы ***точность*** детерминации. Из значения емкости делаем вывод, что из числа студентов, обладающих свойством У, 24% обладают свойством С. Емкость выражает в дополнение к интенсивности как бы ***полноту*** детерминации.

Интенсивность и емкость обладают свойствами, на основе которых достаточно легко интерпретировать детерминацию. Ниже предлагается примерная схема совместной интерпретации значений этих характеристик детерминации.

Интенсивность	I≈0	I=1	I≈0	I≈1
Емкость	C≈0	C≈0	C=1	C≈1
Детерминация	Неполная и неточная	Точная, но неполная	Неточная, но полная	Точная, полная

Схема 3.4.3. Интерпретация детерминации

Еще один способ анализа таблицы (2×2)

О статистической зависимости можно судить по таблицам, приведенным на

схеме 3.4.2, и без использования коэффициентов. Ведь относительные частоты в долях (частости) являются оценками вероятности некоторых событий. Например, обозначим через Р(У,П) вероятность события «быть в будущем политологом и одновременно быть удовлетворенным учебой на четыре балла», через Р(У) -вероятность события «быть удовлетворенным на четыре балла», через Р(П) - вероятность события «быть в будущем политологом». Известно, что если два последних события независимы, то Р(У,П)=Р(У)×Р(П). Для нашего примера (см. таблицу 3.4.1) Р(У)=250/1000=0,25; Р(П)=100/1000=0,1; Р(У,П)=30/1000=0,3. Различие небольшое, поэтому события У и П, скорее всего, независимы.

Для локальной связи пригодны и любые другие меры, существующие для таблиц сопряженности любого размера (r×s), т. е. когда число строк в таблице равно r, а число столбцов равно s. Прежде чем перейти к ним, приведем пример использования на практике дихотомических пар понятий: истинное - ложное значение коэффициента связи; непосредственная связь - опосредованная связь.

Непосредственная - опосредованная связь

По таблице 3.4.1 коэффициент Юла показывает скорее на статистическую независимость, чем на статистическую зависимость, так как Q_1=0,14. Социологу может показаться сей статистический факт странным, так как не согласуется с его содержательными гипотезами. Например, из предыдущих исследований могло быть известно, что студенты-политологи в основном удовлетворены учебой. Сомнения социолога будут вполне оправданы, ибо отсутствие ***непосредственной корреляционной связи*** еще не говорит об отсутствии связи вообще. Связь между двумя свойствами может быть опосредована третьим свойством. Маленькое значение коэффициента может быть обусловлено тем, что ***характер связи*** между «быть политологом» и «быть удовлетворенным учебой» различен, например, для юношей и девушек. Таблица 3.4.2 - таблица сопряженности между свойствами «быть политологом» и «иметь четвертую степень удовлетворенности учебой» для девушек, а таблица 3.4.3 соответственно для юношей. Проверьте: сумма частот в ячейках вида (i,j) в этих двух таблицах равна частоте, соответствующей аналогичной ячейке таблицы 3.4.1.

Таблица 3.4.2

Таблица сопряженности для девушек

	Удовлетворенные учебой на «четыре»	**«Остальные»**	**Итого**
Будущие политологи	20	20	40
«Не политологи»	20	500	520
Итого	40	520	560

Таблица 3.4.3

Таблица сопряженности для юношей

	Удовлетворенные учебой на «четыре»	**«Остальные»**	**Итого**
Будущие политологи	10	50	60

«Не политологи»	200	180	380
Итого	210	230	440

Подсчитаем коэффициент Юла для девушек (Q_f) и для юношей (Q_m). Первый будет равен примерно 0,9, а второй равен - 0,7.

$$Q_f = \frac{20 \times 500 - 20 \times 20}{20 \times 500 + 20 \times 20} \approx 0{,}9 \qquad Q_m = \frac{10 \times 180 - 200 \times 50}{10 \times 180 + 200 \times 50} \approx -0{,}7$$

Во-первых, нетрудно заметить, что в том и другом случае скорее наблюдается статистическая зависимость, чем независимость. Во-вторых, в самом деле характер связи для наших подвыборок действительно различен. Для девушек получен следующий результат: либо почти все будущие политологи удовлетворены, либо не политологи по удовлетворенности относятся к «остальным». Для юношей совершенно другой результат, а именно: либо почти все политологи по удовлетворенности «остальные», либо «не политологи» удовлетворены учебой.

По этой причине значение коэффициента Юла, полученное без учета пола студента, и показало отсутствие связи. Такая ситуация для социолога может быть обозначена как ложное отсутствие корреляционной связи, проистекающее из существования опосредованной связи, характер которой диаметрально противоположный на отдельных группах объектов. Этот пример показывает, что конкретные значения коэффициентов интерпретировать необходимо очень осторожно. Графически этот случай иллюстрирует граф, изображенный на рис.3.4.2. Связь между признаками 1 и 6 не наблюдается. В то же время наблюдается связь между признаками 1 и 5, а также между признаками 5 и 6.

Другая ситуация ложных корреляционных связей является более очевидной. Это когда большое значение коэффициента обусловлено не сильной связью между свойствами, а тем, что существование каждого из этих свойств обусловлено одной и той же причиной. Подозрение вызывает треугольник на том же рис. 3.4.2. Интерпретация больших и маленьких значений коэффициентов требует при анализе особого внимания. Этот вывод относится в равной мере ко всем коэффициентам, с которыми работает социолог.

Переходим к рассмотрению коэффициентов для случая таблиц сопряженности вида (r×s). Вернемся к нашей таблице 3.3.1, где r=6, а s=5. Прежде всего, следует отметить, что в соответствие каждой ячейке можно поставить как прямую детерминацию (от профессии к удовлетворенности) с интенсивностью (процент по строке) и емкостью (процент по столбцу), так и обратную (от удовлетворенности к профессии). Дальнейший анализ таблицы проводится по совокупности этих характеристик. Для выделения сильных локальных связей обычно задаются ограничения на значения интенсивности и емкости. По сути, речь идет о ранжировании всех локальных связей. В этом случае не ставится вопрос о взаимосвязи феноменов «будущая профессия» и «удовлетворенность учебой», а ищутся как бы цепочки детерминации, что в дальнейшем может быть использовано для формирования гипотез о ***факторных*** синдромах и ***причинно-следственных*** отношениях. Напомним, что восходящая стратегия анализа и служит для формирования новых гипотез в исследовании.

Упомянутый выше «язык» анализа локальных связей - язык детерминации - достаточно легко переводится и на многомерный случай. Однако к работе [13] следует обращаться, имея определенный уровень математической подготовки.

Меры связи, основанные на χ^2 (хи-квадрат)

Представим себе, как будет выглядеть наша таблица сопряженности в ситуации статистической независимости между феноменами «будущая профессия» и «удовле-

творенность учебой». Нетрудно вспомнить, что при статистической Независимости, например, для частоты в ячейке (1,4) выполняется соотношение:

$\frac{n_{14}}{n_{04}} = \frac{n_{10}}{n_{00}}$, т.е. $n_{14} = \frac{n_{10} n_{04}}{n_{00}}$.

Если теперь записать это в общем виде, т. е. для любой ячейки (i,j), то в случае статистической независимости будет верно соотношение:

$$n_{ij} = \frac{n_{0j} n_{i0}}{n_{00}}$$

Эту частоту, для ее отличия от реальной, можно назвать теоретической и обозначить через n_{ij}. В таблице 3.4.4 приведены наши реальные частоты, взятые из таблицы 3.3.1, и теоретические. Первые из них — в верхнем левом углу ячейки, а вторые — в нижнем правом углу ячейки.

Таблица 3.4.4

Таблица сопряженности: реальные и теоретические частоты

Будущая профессия студента	Степени удовлетворенности учебой					Маргинальные частоты
	1	2	3	4	5	
1. Политолог	14 20	20 30	31 20	30 25	5 5	100
2. Социолог	30 40	40 60	60 40	60 50	10 10	200
3. Культуролог	90 60	90 90	60 60	45 75	15 15	300
4. Филолог	31 20	30 30	19 20	15 25	5 5	100
5. Психолог	8 10	10 15	15 10	15 12,5	2 2,5	50
6. Историк	27 50	110 75	15 50	85 62,5	13 12,5	250
Маргинальные частоты	200	300	200	250	50	N=1000

Является естественным для определения отклонения от статистической независимости воспользоваться разностью между реальными частотами и теоретическими (для случая статистической независимости), т.е. разностью вида n_{ij}-n^t_{ij}. Как и в случае введения формулы для вычисления дисперсии, нам нужны абсолютные значения этой разности, поэтому возводим ее в квадрат. Этот квадрат делим на теоретическую частоту, т. е. как бы нормируем. Тем самым достигается независимость от объема ячейки. Все ячейки становятся равноправными независимо от их объема. Затем суммируем все эти отклонения по всем 30-ти ячейкам таблицы и получаем величину называемую ***хи-квадрат.*** Она выглядит следующим образом:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}}$$

Для нашего примера эта величина вычисляется как сумма тридцати членов:

$$\chi^2 = \frac{(14-20)^2}{20} + \frac{(20-30)^2}{30} + \Lambda\ \Lambda + \frac{(85-62{,}5)^2}{62{,}5} + \frac{(13-12{,}5)^2}{12{,}5} = 125{,}6$$

Эта величина, эта статистика знаменита тем, что имеет закон распределения, который называется законом распределения ***хи-квадрат.*** Поэтому с ее помощью решается много различных задач, проверяются различные ***статистические гипотезы.*** Нас пока интересует только аспект использования величины хи-квадрат для конструирования мер связи. Самой этой величиной как мерой связи неудобно пользоваться, ибо ее значение может быть каким угодно большим и зависит от размера таблицы сопряженности. Различие в коэффициентах, основанных на хи-квадрат, заключается в определенном нормировании величины хи-квадрат. Одним из часто используемых коэффициентов является коэффициент взаимной сопряженности Пирсона. Он имеет следующий вид:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}},$$

где N- общее число объектов. В нашем случае объекты - студенты-гуманитарии. Раньше их число мы обозначали через n_{00}, которое было равно 1000. Для наших целей так было удобнее, а в данном случае нет никакой необходимости ни в двойных индексах, ни в индексах вообще.

Если значение коэффициента получится близким к нулю или равным нулю, то это означает статистическую независимость признаков. Случай близости значения к единице будет говорить о статистической зависимости. Значение коэффициента ни при каких условиях не достигает единицы, но для социолога это не имеет никакого принципиального значения. Для нашей таблицы сопряженности χ^2= 125,6, а значение C=0,33. Опять-таки возникает вопрос о значимости отличия такого значения от нуля.

О значимости значений коэффициентов

Определяются такого рода значимости на основе проверки статистических гипотез. Эти гипотезы не следует путать с так называемыми содержательными гипотезами исследования. Разумеется, в ряде случаев гипотеза исследования может быть сформулирована и в виде статистической гипотезы. Проверка статистической гипотезы о значимости отличия значения коэффициента от нуля возможна при условии существования ***закона распределения*** коэффициента.

Что это означает? Предположим, каждый из вас для изучения студентов-гуманитариев (это наша генеральная совокупность) сформировал «отличную» выборку и подсчитал значение, например, коэффициента Юла. Какими бы «хорошими» ни были выборки, на каждой из них будет получено свое собственное значение этого коэффициента. Совокупность таких значений подчиняется и может быть описана некоторым законом распределения. Для коэффициента Юла известно, что он имеет вполне ***определенный*** закон распределения. Если для коэффициента теоретический закон распределения известен, то такой коэффициент называется ***статистикой*** в отличие от эвристики. Не надо путать с тем, что статистикой называют и просто совокупность данных в той области науки, которая называется ***статистикой.*** Мы сейчас рассуждаем в рамках другой науки, которая называется ***математической статистикой.***

Каждый закон распределения имеет ***параметры.*** Примером закона является уравнение прямой $y = aX + b$. Это семейство прямых. Здесь параметрами являются a, b. Аналогично можно рассуждать во всех случаях законов, известных вам из школьной программы (парабола, гипербола, синусоида и т. д.). Только теперь вы имеете дело с более сложными законами: нормальным, хи-квадрат и т. д. Более того, для некоторых законов, например для хи-квадрат, даже нельзя в явной форме записать формулу.

Некоторые законы ***табулированы,*** т. е. существуют математические таблицы (они есть во многих книгах, где описываются методы математической статистики), из которых можно определить табличное значение некоторой статистики при заданных параметрах распределения. Например, табличное значение для величины «хи-квадрат» - это то значение, которое оно принимает при статистической независимости.

Кроме параметров для обращения к математическим таблицам необходимо обязательно задать так называемый ***уровень значимости*** *(а*), т. е. уровень возможной ошибки. В математической статистике на основе данных выборки ни один вывод не делается без некоторой ошибки. Значение *а* может быть равным 0,10; 0,05; 0,01. Тогда наши выводы будут верны в 90 случаях из ста, если социолог задал первое из этих значений. Для второго уровня значимости выводы верны в 95 случаях из ста, а для третьего - в 99 случаях из ста, а для четвертого 999 случаев из тысячи.

Таким образом, если некоторая величина табулирована, то, задавшись уровнем значимости и параметрами закона распределения, можно узнать ее теоретическое значение. А у нас всегда есть реальное значение. Сравнение этих значений и позволяет проверять статистические гипотезы.

Возвращаясь к коэффициенту Юла и статистики «хи-квадрат», следует сказать, что первый из них имеет нормальный закон распределения, а второй - распределение хи-квадрат. Параметром для нормального закона является дисперсия, а параметром для хи-квадрат - число степеней свободы, равное (r-l)(s-l). По существу, число степеней свободы - число ячеек в таблице сопряженности, которые могут изменяться свободно (отсюда и название число «степеней: свободы») при заданных маргинальных частотах. В нашем случае реальное значение «хи-квадрат» равно χ^2=125,6, а табличное значение χ^2_t=10,85 при уровне значимости, равной 0,05, и числе степеней свободы (r-l)(s-l)=20. Таким образом, $\chi^2 > \chi^2_t$, т. е. отклонение от нуля значимо. Признаки «будущая профессия студента» и «удовлетворенность учебой» статистически зависимы.

Понятие значимости тесно связано с понятием «доверительный интервал». Для каждой статистики это интервал, в котором содержится «истинное» (для генеральной совокупности) значение этой статистики. Если истинное значение коэффициента Юла обозначить через Q^t, а реально вычисленное через Q, то доверительный интервал выглядит:

$$\mathbf{Q - \Delta \le Q^t \le Q + \Delta.}$$

Для каждой статистики величина Δ определяется в зависимости от закона распределения статистики и, естественно, с помощью математических таблиц, где эти законы табулированы. Приводить формулы для вычисления доверительных интервалов мы не будем. К примеру, социолога всегда интересует значимость процентов. В работе [8, с. 191—195] вы можете найти формулу для вычисления доверительного интервала в этом случае.

Из такого упрощенного анализа значимости и законов распределения социологу необходимо усвоить, что умные люди, работающие в далекой от него науке под названием математическая статистика, владеют большим аппаратом для решения социологических задач. Это не означает, что вы должны эту науку изучить досконально, но это означает, что Вы должны научиться задавать таким людям правильно поставленные вопросы и не ожидать от математики того, чего она не может дать.

Задание на семинар или для самостоятельного выполнения

Каждому студенту на основе своей собственной таблицы сопряженности необходимо выполнить следующие задания:

1. Обозначить одну из градаций (любого из двух признаков таблицы сопряженности) как целевое свойство. Подсчитать значения коэффициента Юла между этим целевым признаком и несколькими другими. Провести ранжирование полученных значений по степени их влияния на целевой признак.

2. Вычислить интенсивность и емкость детерминации для нескольких свойств и на основе сравнения сделать соответствующие выводы.

3. Вычислить значение хи-квадрат и сравнить с табличным при различных уровнях значимости. Сделать соответствующие выводы.

5. МЕРЫ СВЯЗИ: ОСНОВАННЫЕ НА МОДЕЛИ ПРОГНОЗА И РАНГОВЫЕ

Модальные меры Гуттмана. Сравнение распределений посредством меры Л. Гудмена и Е. Краскала. Когда социолог имеет дело с ранжированными рядами? Принцип сравнения ранжированных рядов. Связанные ранги. Коэффициенты ранговой корреляции Л. Гудмена и Е. Краскала, Р. Somерса, М. Дж. Кендалла.

Вначале мы приведем примеры коэффициентов связи для признаков, имеющих по-прежнему номинальный уровень измерения. Особое внимание к такого рода мерам вполне оправданно. Специфика социологических данных такова, что социолог в основном работает с номинальным уровнем измерения. Исключение составляют первый (государственная статистика) и третий (бюджеты времени) типы социологической информации. Как и раньше, в качестве примера рассматриваем связь между будущей профессией студента и удовлетворенностью учебой. Это несмотря на то, что второй из них измерен по порядковой шкале. Пока эту упорядоченность никак не используем. Социологу приходится часто так поступать, ибо он всегда работает с эмпирией в ситуации разнотипности шкал. Мер, учитывающих эту разнотипность, мало, и они не всегда удовлетворяют потребностям социолога. В силу этого приходится намеренно идти на «огрубление» данных и работать в ситуации номинального уровня измерения даже тогда, когда речь идет о порядковых и «метрических» шкалах. *Следует вас предостеречь. Во многих работах, упомянутых в списке литературы, содержатся разного рода неточности и некоторые ошибки в написании формул. Поэтому при самостоятельном изучении следует перепроверять формулы, сравнивая их с аналогичными из других источников.*

Прежде всего, рассмотрим меры, основанные на так называемой модели прогноза. Это уже как бы другой «язык» анализа таблиц сопряженности. Для социолога понятие «прогноз» носит не только многозначный характер, но к этому понятию отношение очень осторожное и трепетное. Если на основе эмпирических данных и можно что-то прогнозировать, предсказывать, то в достаточно узком смысле понимания прогноза. При этом ход рассуждений примерно такой. Если ничего не изменится, то может быть то-то и то-то. Социологи-математики (такие тоже есть) термин «прогноз, предсказание» употребляют в еще более узком смысле, но очень часто [4, 5]. Мы также будем пользоваться понятием «прогноз» в очень узком смысле. Попробуем коротко и грубо прояснить, в каком смысле.

У нас с вами есть одномерное распределение какого-то признака. Напоминаем, что под признаком понимаем как отдельно взятый эмпирический индикатор (наблюдаемый признак), так и производный от эмпирических индикаторов показатель. Пусть таковым признаком будет удовлетворенность учебой (У). Распределение этого признака можем интерпретировать следующим образом. Есть значения признака (различные степени удовлетворенности учебой), и есть вероятности этих значений (относительные частоты в долях или частости). А, точнее, оценки вероятности, полученные по выборке. Все, что рассчитывается по выборочной совокупности, называет-

ся ***оценками истинных*** (существующих для изучаемой генеральной совокупности) ***значений.*** Разумеется, социолог может опускать термин «оценка», если понимает, о чем идет речь. Для простоты мы будем поступать так же.

Итак, наши вероятности P_{0j} равны маргинальным частотам по столбцам (именно они соответствуют признаку (У) - удовлетворенность учебой), деленным на общее число опрошенных студентов-гуманитариев (n_{00}). В виде формулы это выглядит так:

$$P_{0j} = \frac{n_{0j}}{n_{00}}.$$

Тогда, по приведенной ниже таблице 3.5.1 (это та же таблица сопряженности, с которой мы постоянно работаем), вероятности пяти степеней удовлетворенности учебой равны:

$$P_{01} = 200/1000 = 0{,}2;$$
$$P_{02} = 300/1000 = 0{,}3;$$
$$P_{03} = 200/1000 = 0{,}2;$$
$$P_{04} = 250/1000 = 0{,}25;$$
$$P_{05} = 50/1000 = 0{,}5$$

Эти вероятности можно интерпретировать как вероятности ***статистического предсказания*** *(У).* Мы же их получили по «хорошей» выборке. Поэтому если из нашей изучаемой генеральной совокупности студентов-гуманитариев случайно выберем некоторого студента, то вероятность того, что у этого случайного студента окажется максимальная удовлетворенность учебой, очень мала. Это потому, что по выборке она была равна всего лишь 0,05. Вероятность «отгадать» все остальные варианты удовлетворенности учебой тоже невелика, ибо они, как видите, не больше, чем 0,3. При этом само понятие «вероятность» можно трактовать на уровне обыденного сознания. Только в повседневной жизни вам обычно говорят, например, «вероятность того, что у меня завтра будет плохое настроение для прогулки, равна 90%» или «вероятность того, что я завтра приду к тебе в гости, меньше 50%» или «вероятность нашей возможной встречи «фифти - фифти» (50 на 50)». И вы всегда понимаете, что сие означает. При этом такие суждения вы интерпретируете не столько количественно, сколько качественно. А в математических формулах пользуются не процентами для оценки вероятности, а долями - частостями — и, соответственно, вероятность принимает вполне конкретное значение из интервала от 0 до 1.

Теперь вполне правомерно поставить вопрос: **Как изменятся рассчитанные нами вероятности иметь ту или иную степень удовлетворенности учебой, если привлечь к анализу второй признак** (будущую профессию студента)? Можно вопрос поставить и по-другому: **Насколько знание будущей профессии прибавит знания об удовлетворенности учебой?** Или: **Насколько информация о будущей профессии изменит информацию об удовлетворенности учебой*?*** Поиск ответа на последний вопрос порождает меры связи, основанные на понятии ***энтропии*** (мы касались этого понятия при введении качественных коэффициентов вариации). Такого рода меры мы не будем рассматривать. Вы можете с ними познакомиться в работах [3, 8, 11].

Первый наш вопрос можно поставить и так: ***Как и насколько изменятся вероятности предсказания удовлетворенности учебой, если учесть будущую профессию?*** Как вы уже догадываетесь, по сути, речь идет о знании условных распределений нашего признака (У) или условных частот, или условных вероятностей, т, е. вероятностей, которые логично обозначить как $P_{j/i}$. Индекс первый (j) относится к столбцам, т. е. к удовлетворенности учебой (признак У), второй (i) относится к строкам, т. е. к будущей профессии (признак Х), а косая черта подчеркивает, что признак

(X) является условием.

Существуют всевозможные коэффициенты, помогающие найти ответ на подобные вопросы. Как видно из наших рассуждений, они должны быть ***направленными*** и носить, так же как и меры, основанные на хи-квадрат, характер ***«глобальный»,*** т. е. давать оценку связи в целом для всей таблицы сопряженности в отличие от локальных мер (связь отдельных свойств).

Если для кого-то термин «предсказание» остался пока непонятым, то при описании предлагаемых ниже мер как можно реже будем пользоваться этим термином.

Меры λ (лямбда) Л. Гуттмана (L. Guttmann)

Таких мер три, две из них направленные, а одна представляет собой усреднение первых двух. Мы приведем только одну $\lambda_{y/x}$. Эта мера, этот коэффициент характеризует в случае нашей задачи влияние будущей профессии (X) на удовлетворенность учебой (У). Отвечает на вопрос, насколько изменяется предсказание (У) при знании (X). Ниже приводится формула, в которой используются известные вам обозначения, за исключением:

n_{imax} - максимальная частота в i-й строке:

n_{0max} - максимальная частота среди маргинальных частот по столбцам.

$$\lambda_{y/x} = \frac{\sum_{(i)} n_{i\max} - n_{0\max}}{n_{00} - n_{0\max}}$$

Эта формула была бы понятнее, если вместо частот использовать частости (доли), интерпретируемые как вероятности [11, с. 126], Такую формулу мы не будем приводить, чтобы не пугать излишними формулами. Отметим лишь, что в литературе приводится как формула, записанная через абсолютные частоты, так и через частости. Кроме того, фамилия Гуттмана тоже приводится по-разному. Например, Гудман в работе [8, с. 131]. Это не так уж важно.

Для того чтобы пояснить содержательный смысл этой меры, этого коэффициента, ниже приводится та же таблица сопряженности, с которой мы постоянно работаем для изучения взаимосвязи между «будущей профессией студента» (признак X) и «удовлетворенностью учебой» (признак У). Таблица 3.5.1 содержит те же частоты, что и таблица 3.3.1, за исключением обозначений самих частот. В нее добавлен новый столбец - последний с максимальными частотами по всем строкам, включая строку с маргинальными частотами по столбцам. Они нам необходимы для вычисления коэффициента $\lambda_{y/x}$ Гуттмана.

Таблица 3.5.1

Таблица сопряженности двух признаков (У) и (X).

Будущая профессия студента (X)	**Степень удовлетворенности учебой (У)**					**Маргинальные частоты по строкам n_{i0}**	**Максимальные частоты по строкам $n_{i\ max}$**
	1	**2**	**3**	**4**	**5**		
1. Политолог	14	20	31	30	5	100	$n_{1\ max}$=31
2. Социолог	30	40	60	60	10	200	$n_{2\ max}$=60
3. Культуролог	90	90	60	45	15	300	$n_{3\ max}$=90
4. Филолог	31	30	19	15	5	100	$n_{4\ max}$=31
5. Психолог	8	10	15	15	2	50	$n_{5\ max}$=15
6. Историк	27	110	15	85	13	250	$n_{6\ max}$=110
Маргинальные частоты по	200	300	200	250	50	n_{00}=1000	$n_{0\ max}$=300

столбцам n_{0j}							

Чему же равен коэффициент в нашем случае? Он рассчитывается очень просто.

$$\lambda_{y/x} = \frac{(31+60+90+31+15+110)-300}{1000-300} = 0{,}05$$

Даже по тому, как вычисляется коэффициент, видно, что он позволяет определять, существуют ли в строках модальные группы, т. е. есть ли в каждой профессиональной группе ярко выраженная, часто встречаемая «степень удовлетворенности учебой». Суда по нашей таблице, таких групп практически нет, что и подтверждается маленьким значением коэффициента. Какими же свойствами обладает этот коэффициент?

1. Он изменяется от нуля до единицы.

2. Он равен единице только в одном случае, когда в каждой профессиональной группе все студенты имеют одинаковую степень удовлетворенности учебой и при этом в каждой отличную от другой. Если бы наша таблица сопряженности при тех же маргинальных частотах имела бы такой вид, как это представлено в таблице 3.5.2, коэффициент был бы равен 0,86.

$$\lambda_{y/x} = \frac{(100+200+300+100+50+150)-300}{1000-300} = \frac{900-300}{700} = 0{,}86$$

Таблица 3.5.2

Таблица сопряженности двух признаков (для $\lambda_{y/x} = 0{,}86$)

Будущая профессия (X)	**Степени удовлетворенности (У)**					**Маргинальные частоты по строкам** n_{i0}	**Максимальные частоты по строкам** n_{imax}
	1	2	3	4	5		
1.Политолог	0	0	100	0	0	100	n_{1max}=100
2.Социолог	200	0	0	0	0	200	n_{2max}=200
3.Культуролог	0	300	0	0	0	300	n_{3max}=300
4.Филолог	0	0	0	100	0	100	n_{4max}=100
5.Психолог	0	0	0	0	50	50	n_{5max}=50
6.Историк	0	0	100	150	0	250	n_{6max}=150
Маргинальные частоты по столбцам n_{0i}	200	300	200	250	50	n_{00}=1000	n_{0max}=300

Итак, визуально мы наблюдаем наличие модальных групп в строках, кроме последней. Если бы в нашей таблице число строк равнялось числу столбцов, например, не было бы историков, то коэффициент был бы равен 1, а таблицу можно было бы перестановкой столбцов превратить в такую, в которой только диагональные элементы отличались бы от нуля. Таким образом, по значению коэффициента можно судить о степени отличия реальной таблицы от диагональной. В случае, когда значение коэффициента равно 1, вероятность ***статистического предсказания*** (У) по Х максимальная. Такой случай практически в социологических исследованиях не встречается.

3. Значение коэффициента равно нулю в нескольких случаях. Первый - все частоты сосредоточены только в одной строке. На самом деле знание признака Х нечего не дает для увеличения знания об У. Второй случай - отсутствие феномена модальности, т. е., условно говоря, полная «размытость» данных в таблице. По таблице 3.5.1 мы получили значение, близкое к нулю и равное 0,05. Практически модальность не наблюдается. И наконец; третий случай, когда все частоты сосредоточены только в одном столбце.

Этот случай заслуживает особого внимания, ибо противоречит основному содержанию коэффициента. Если данные сосредоточены в одном столбце, то естественно модальные классы существуют. Тогда и вероятность предсказания значения У по значению Х должна быть равна единице. А наш коэффициент равен нулю. Здесь мы наблюдаем ситуацию, когда ***коэффициент плохо ведет себя в нуле.*** Запомните эту фразу. Вы будете встречаться с подобными фразами и в случае других коэффициентов. Чтобы исключить неверную интерпретацию нулевого значения, необходимо по одномерному распределению уточнить, не сосредоточены ли данные только в одном столбце. Такой случай также не встречается в социологической практике.

Представляется важным отметить, что в реальных исследованиях значения коэффициента Гуттмана очень малы, и использовать их нужно так же, как и многие другие коэффициенты в сравнительном контексте, например, для ранжирования как бы независимых между собой признаков по степени их влияния на некоторый особенно важный для исследователя признак, обозначаемый как целевой, зависимый. Если такого нет, то направленные коэффициенты «лямбда» использовать не имеет особого смысла.

Меры τ (тау) Л. Гудмена и Е. Краскала (L. Goodman, E. Kruskal)

Эти меры, на мой взгляд, интересны социологу, ибо с ними можно работать в сравнительном контексте, не обращая особого внимания на всякие ***значимости.*** Таких мер вообще-то три, как и в случае мер Гуттмана. Первые две из них направленные, а третья как бы усредняет первые два. Мы рассмотрим только одну из них. Для этого опять обратимся к нашей таблице сопряженности 3.5.1. При этом вспомним и рис. 3.3.1. На этом рисунке были изображены эмпирические кривые распределения удовлетворенности учебой в каждой профессиональной группе -будущие профессии студентов-гуманитариев (мы уже обозначили эти признаки через У и Х). Визуально мы *с* вами наблюдали наличие трех ***типологических синдромов*** по характеру распределения признака У. Другими словами, три типа структуры удовлетворенности учебой.

Ни один коэффициент глобального характера не позволит определить, сколько типов структур наблюдается. Если социолога интересуют такие группы, то до применения всяких коэффициентов представляется целесообразным хотя бы визуально на компьютере просмотреть графики такого вида, которые изображены на рис. 3.3.1 и рис. 3.3.2. Тот же коэффициент, который мы рассмотрим, позволяет ***в целом*** определить степень отличия условных распределений У от безусловного. Ниже приведем формулу. В ней будем использовать обозначения вероятностей (условных и безус-

ловных), введенных в начале этого раздела. В этот раз формулу запишем не на языке абсолютных частот, а на языке вероятности - доли, частости. В литературе она приводится обычно через абсолютные частоты [1, с. 36, 3, с. 36].

Один из трех коэффициентов *τ (тау) Гудмена и Краскала* выглядит следующим образом.

$$\tau_{y/x} = \frac{\sum_i p_{io} \sum_j (p_{j/i} - p_{oj})^2}{1 - \sum_j p_{oj}^2}.$$

Если вы подставите в эту формулу вместо вероятности (точнее оценок вероятности) частоты, то получите формулу, приводимую в литературе, т. е.:

$$p_{oj} = \frac{n_{oj}}{n_{oo}}; p_{io} = \frac{n_{io}}{n_{oo}}; p_{j/i} = \frac{n_{ij}}{n_{io}}.$$

Две первые формулы служат для вычисления безусловных вероятностей. Их значения приведены соответственно в последней строке таблицы 3.5.3 и в последнем столбце. Третья формула — для вычисления условной вероятности. Значения такой вероятности приведены в ячейках таблицы 3.5.3. Они аналогичны данным таблицы 3.3.2 (верхнее левое значение в ячейках).

Таблица 3.5.3

Таблица сопряженности (условные и безусловные вероятности)

Будущая профессия студента	**Степень удовлетворенностью учебой**					**Безусловные вероятности P_m**
	1	2	3	4	5	
1.Политолог	0,14	0,20	0,31	0,30	0,05	0,10
2. Социолог	0,15	0,20	0,30	0,30	0,05	0,20
3.Культуролог	0,30	0,30	0,20	0,15	0,05	0,30
4. Филолог	0,31	0,30	0,19	0,15	0,05	0,10
5. Психолог	0,16	0,20	0,30	0,30	0,04	0,05
6. Историк	0,11	0,44	0,06	0,34	0,05	0,25
Безусловные вероятности P_m	0,20	0,30	0,20	0,25	0,05	N = 1000

Коэффициент «τ» чем-то напоминает и «**хи-**квадрат», и λ Гугтмана. Однако он не такой «прозрачный» для объяснения, как эти коэффициенты. Вообще-то говоря, если все можно было бы описывать и объяснять в социологии вербально, то, может, язык математики был бы и не нужен. И что совершенно очевидно, чем ближе язык математики к языку социолога, тем он сложнее. Все-таки попытаемся прояснить содержательный смысл приведенного коэффициента.

Прежде всего, необходимо пояснить, зачем при сравнении распределений всякие квадраты. В числителе квадрат по аналогии с формулой дисперсии. Для того чтобы учесть отклонение условной частоты от безусловной в одну и другую сторону. В знаменателе сумма квадратов безусловных вероятностей. Простая их сумма всегда равна единице. Это вы знаете. Такой знаменатель - количественная характеристика распределения по столбцам (безусловное распределение по У). Числитель несет в себе основное содержание коэффициента. В числителе в скобках - отклонение условной вероятности от безусловной вероятности У. Естественно, все отклонения суммируются по всем значениям У (по всем столбцам). В свою очередь такие величины, полученные по каждой строке (по каждому условному распределению У) суммируются как бы с весами, равными безусловной вероятности по строке. Тем самым строки уравниваются в «правах» за вклад в значение коэффициента. Напомню, что при вычислении величины «хи-квадрат» мы уравнивали в «правах» ячейки таблицы сопряженности, а здесь - строки.

Коэффициент *τ (may)* Гудмена и Краскала обладает следующими свойствами:

1. Принимает значение от нуля до единицы.

2. Равен нулю, если структура распределения по строкам одинакова и такая, как структура распределения маргинальных (по столбцам) частот. В этом случае наблюдается ***статистическая независимость*** У от Х. Будущая профессия не влияет на удовлетворенность учебой.

3. Равен единице, если будущая профессия студента полностью детерминирует его удовлетворенность учебой. Каждой профессии соответствует своя собственная степень удовлетворенности учебой. Чисто формально это означает, что таблицу сопряженности можно привести к диагональному виду. В самом деле, для таблицы 3.5.2 значение коэффициента равно $\tau_{y/x}$=0.83

Вычислим значение коэффициента для нашей таблицы 3.5.3. Чтобы вычислить числитель, нужно сложить 6 (для всех строк таблицы) величин. Каждая такая величина равна

$$p_{io}\sum_{j}(p_{j/i}-p_{jo})^2,$$

Для первой строки она равна:

$0{,}1\{(0{,}14\text{-}0{,}20)^2+(0{,}20\text{-}0{,}30)^2+(0{,}31\text{-}0{,}20)^2+(0{,}30\text{-}0{,}25)^2+ +(0{,}05\text{-}0{,}05)^2\} = 0{,}0028$

Для остальных строк эта величина соответственно равна 0,0045; 0,006; 0,0022; 0,00121; 0,01385. Таким образом, значение числителя равно 0,024. Знаменатель равен:

$1-\{(0.2)^2+(0.3)^2+(0.2)^2+(0.25)^2+(0.05)^2\} = 0.77$

Тогда значение коэффициента будет равно $\tau_{y/x}$=**0.03**. Такое небольшое значение коэффициента говорит об отсутствии влияния будущей профессии на структуру удовлетворенностью учебой. Вероятность предсказания удовлетворенности учебой практически не изменится, если учитывать будущую профессию.

До сих пор мы с вами рассматривали только меры связи для номинальных признаков, ибо они чаще других встречаются в социологических данных. При этом, ана-

лизируя данные нашей таблицы сопряженности, мы не обращали внимания на то, что один из признаков имел порядковый уровень измерения. Не использовать информацию об упорядоченности - значит намеренно отказаться от ценной информации. Разумеется, существуют коэффициенты, позволяющие учесть то, что один из сопрягаемых признаков измерен по порядковой шкале.

Существует так называемый ранговый бисериалъный коэффициент для случая изучения связи между дихотомическим (поэтому коэффициент называется бисериальным) номинальным признаком иранговым [2, с. 165—167, (8, с. *139, 11, с. 121]. При этом для случая несвязанных рангов. Напомним, что с ситуацией связанных рангов мы встречаемся, если в ранжированном ряду есть одинаковые ранги. Также существует точечный бисериалъный коэффициент для случая изучения связи между дихотомическим номинальным признаком и «метрическим».*

Ранговые коэффициенты связи

Ранговыми коэффициентами связи называются меры связи, позволяющие вычислять степень согласованности в ранжировании одних и тех же объектов по двум различным основаниям или по двум различным признакам. Мы неоднократно ссылались на необходимость для социолога такого рода коэффициентов. Например, при построении шкалы суммарных оценок появлялась необходимость в проверке согласованности результатов, полученных по итоговой шкале, с данными по исходным шкалам (суждениям).

Коэффициентов ранговой корреляции много. Для того чтобы понять их схожесть и различие, необходимо вначале несколько отойти от таблиц сопряженности и нашей задачи. А вам придется вернуться к разделу книги, посвященному процедуре ранжирования. Как было отмечено, такая процедура возникает у социолога как на этапе измерения, так и на этапе анализа данных. В любом случае возникает задача определения степени согласованности двух ранжированных рядов. Представим себе, что для одной и той же совокупности объектов получили два ранжированных ряда. Например, по тем же будущим профессиям студента. Значит, объектов у нас всего шесть по числу профессий. Пусть первый ряд получен по степени уменьшения индекса удовлетворенности учебой. Второй ряд - по степени уменьшения индекса уверенности в трудоустройстве по профессии после окончания вуза. Далее будем коротко называть эти признаки - «удовлетворенность» и «уверенность».

В данном контексте мы не будем обсуждать вопрос, каким образом измерены эти признаки как характеристики группы. Заметим лишь, что они могли быть получены с помощью шкалы суммарных оценок или как групповые индексы, примеры которых были приведены в «Лекциях».

В случае полной (максимальной) согласованности ранжирования по этим двум признакам естественно предположить наличие тесной (сильной) связи между признаками «удовлетворенность» и «уверенность». Такая связь может быть и ***прямой*** (чем больше удовлетворенность, тем больше уверенность), и ***обратной*** (чем больше удовлетворенность, тем меньше уверенность). Из этого проистекает, что логично изменяться значениям коэффициента ранговой корреляции от -1 до +1, Этим свойством обладают все приведенные ниже коэффициенты.

Приведем примеры нескольких коэффициентов, а затем поясним их содержательный смысл,

Мера γ (гамма) Л. Гудмена и Е. Краскала (L.Goodman, E.Kruskal)

$$\gamma = \frac{S - D}{S + D}$$

Мера τ_k (тау) М. Дж. Кендалла (M.Kendall)

$$\tau_k = \frac{(S - D)}{\sqrt{(S + D + T_y)(S + D + T_x)}}.$$

Меры d Р. Сомерса (R.Comers)

$$d_{y/x} = \frac{S - D}{S + D + T_y}.$$

Первая из этих мер в работе [8, с. 135], обозначена как «γ Гудмана». Эти меры удачно описаны в работе [1, с. 37—40]. Вы, конечно, обратили внимание, что у всех приведенных мер один и тот же числитель, а знаменатели различны. Прежде всего, рассмотрим числитель, ибо он несет в себе основное содержание коэффициентов. В таблице 3.5.4 представлены два ранжированных ряда. Объекты ранжирования - будущие профессии. Они приведены в таблице для удобства в том порядке, в котором их ранги во втором ряду возрастают, т. е. в порядке убывания степени уверенности. Число рангов равно числу объектов, ***связанных*** рангов (одинаковых) в наших рядах не наблюдается.

Таблица 3.5.4

Примеры двух ранжированных рядов

Признаки	**Объекты ранжирования**					
	социологи	психологи	политологи	культурологи	историки	филологи
x Удовлетворенность	3	4	2	6	1	5
y Уверенность	1	2	3	4	5	6

Из этой таблицы видим, что политологи в первом ряду имеют ранг 2, а во втором - ранг 3, а историки в первом ряду - ранг 1, во втором - ранг 5. Для того чтобы оценить степень согласованности наших, грубо говоря, «ранжировок», можно применить тот же прием, который был применен при вычислении меры качественной вариации. Образуем из наших шести объектов различные пары. Таких пар будет 6×5/2=15. Возьмем отдельную пару объектов. Ранги, соответствующие первому объекту, обозначим ($\mathbf{i_1, j_1}$), а второму - ($\mathbf{i_2, j_2}$).

Эти ранги могут находиться в различных отношениях. Возможна одна из двух ситуаций, каждая из которых включает два возможных соотношения между рангами (1а, 1б, 2а, 2б).

Первая ситуация:

1а. $\mathbf{i_1 > i_2}$ **и** $\mathbf{j_1 > j_2}$

или соотношение

1б. $\mathbf{i_1 < i_2}$ **и** $\mathbf{j_1 < j_2}$

Вторая ситуация:

2а. $\mathbf{i_1 > i_2}$ **и** $\mathbf{j_1 < j_2}$

или соотношение

2б. $\mathbf{i_1 < i_2}$ **и** $\mathbf{j_1 > j_2}$

В первой ситуации ранги как бы согласованы, а во втором не согласованы. Подсчитаем, для скольких пар из 15-ти наблюдается согласованность, и обозначим число таких пар через S. Затем подсчитаем, для скольких пар наблюдается несогласованность, и обозначим число таких пар через D. В числителе всех приведенных выше мер стоит как раз разница между числом согласованных и несогласованных пар объектов. Для примера наших ранжированных рядов величина (S-D) равна:

S-D = (3-2) + (2-2) + (2-1) + (0-2) + (1-0) = 1.

Здесь первая скобка - результат анализа согласованности / несогласованности рангов в парах, образованных первым объектом с остальными пятью, т. е. в парах (1 и 2), (1 и 3), (1 и 4),(1 и 5), (1 и 6). Среди них согласованность (случай 1а) — в трех парах, а несогласованность (случай 2б) — в двух парах. Вторая скобка - результат анализа пар, образованных вторым объектом, т. е. пар (2 и 3), (2 и 4), (2 и 5), (2 и 6). Среди них в двух парах согласованность, а в двух - несогласованность. Последняя скобка - результат анализа пары (5 и 6).

Мы рассматривали случай отсутствия связанных рангов, поэтому для определения степени согласованности можно использовать первый из трех коэффициентов, приведенных выше. Знаменатель для его вычисления равен:

S+D = (3+2) + (2+2) + (2+1) + (0+2) + (1+0) = 15

или просто числу различных возможных пар, т. е.

6×5/2=15

Тогда $\gamma \approx 0{,}07$. В самом деле степень согласованности в наших ранжированных рядах очень мала. Второй из трех коэффициентов учитывает наличие связанных рангов. Кроме соотношений (1а; 16; 2а; 2б) при анализе пар могут встретиться и другие соотношения (в случае *связанных рангов):*

Третья ситуация:

3а. $\mathbf{i_1 > i_2}$ **и** $\mathbf{j_1 = j_2}$

или

3б. $\mathbf{i_1 < i_2}$ **и** $\mathbf{j_1 = j_2}$

Четвертая ситуация:

4а. $\mathbf{i_1 = i_2}$ **и** $\mathbf{j_1 < j_2}$

или

4б. $\mathbf{i_1 = i_2}$ **и** $\mathbf{j_1 > j_2}$

Число пар, соответствующих третьей ситуации (есть связанные ранги во втором ряду), обозначим через T_y. Число пар, соответствующих четвертой ситуации (есть связанные ранги в первом ряду), обозначим через T_x. Второй коэффициент учитывает число связанных рангов в том и другом ранжированных рядах.

И, наконец, обратите внимание на коэффициент $d_{y/x}$. Мер Сомерса всего три по аналогии с мерами «лямбда» Гуттмана и «гамма» Гудмена и Краскала, т. е. ранговые коэффициенты связи бывают и ***направленные***. Мы привели только одну из трех мер Сомерса. В случае ее использования вопрос о степени согласованности в ранжированных рядах звучит несколько иначе, а именно: влияет ли «уверенность» на «удовлетворенность» и, наоборот, влияет ли ранжирование по «удовлетворенности» на ранжирование по «уверенности». Разумеется, только в смысле того, что ранжирование объектов по степени убывания «уверенности» (признак У) зависит от ранжирования по степени убывания «удовлетворенности» (признак Х). Поэтому в знаменателе учитываются связанные ранги только для признака У.

А теперь представим себе, что речь идет об анализе связи по таблице сопряженности (корреляционная таблица) двух признаков, имеющих порядковый уровень измерения. Допустим, что у каждого нашего студента-гуманитария есть оценка не только удовлетворенности учебой, но и удовлетворенности собой. Оба признака имеют порядковый уровень измерения. Для изучения связи между ними используются те же ранговые меры связи. Их значения рассчитываются по тем же формулам, ибо можно всех наших студентов (объекты ранжирования) упорядочить и получить два ранжированных ряда. Первый - по степени убывания (возрастания) удовлетворенности учебой, а второй - по убыванию (возрастанию) удовлетворенности собой. Естественно, у нас будут сплошь связанные ранги. Напомним, что число рангов равно числу объектов, т. е. 1000. Реально никто такое ранжирование не проводит, а просто вычисляются по таблице сопряженности число согласованных пар, число несогласованных и число связанных рангов. Существуют коэффициенты ранговой корреляции для быстрого счета (коэффициент Спирмена), но в век компьютеров они уже утратили свою актуальность.

Мы рассмотрели все коэффициенты, необходимые для первоначального понимания того, что они из себя представляют, и почему их так много. В завершение этого раздела книги несколько слов о том, что все эти коэффициенты являются статистиками, т.е. для них можно построить доверительный интервал. Тот интервал, в котором находится истинное значение коэффициента, т. е. для изучаемой генеральной совокупности. Доверительные интервалы есть для «лямбда» [1, с. 34], «тау» [1, с. 36], для коэффициентов ранговой корреляции [9, с. 185—187].

В рамках книга не ставилась цель привести все меры или дать их классификацию, ибо для этого необходимы серьезные знания в области науки под названием теория вероятности и математическая статистка. Более того, мы намеренно не рассматривали меры для изучения связи между признаками, измеренными по «метрическим» шкалам (по всем, по которым уровень измерения выше порядкового). Такая позиция обусловлена сочетанием двух факторов процесса обучения студентов. Во-первых, в эмпирической социологии такого рода шкалы встречаются реже других. Во-вторых, в читаемом студентам курсе «Теория вероятности и математическая статистика» понятие «связь» вводится именно с такого рода мер связи.

Задание на семинар или для самостоятельного выполнения

Задание выполняется индивидуально. Каждый студент работает с той же матрицей данных (см. первое задание в начале этой главы), с той же таблицей сопряженности.

1. Вычислить значения направленных мер связи Гуттмана, т. е. вычислить два значения. Сравнить результаты с аналогичными результатами других студентов.

2. Вычислить значения двух направленных коэффициентов Гудмена и Краскала. Сравнить со значениями, полученными в предыдущем задании.

3. Получить два ранжированных ряда. Объектами ранжирования будут группы, полученные при различных значениях первого признака (номинальный уровень измерения). В каждой группе подсчитать среднее арифметическое значение третьего признака (метрический уровень измерения) и упорядочить эти группы в порядке убывания / возрастания этих значений. Тем самым получается **первый** ряд. Для получения второго ряда в тех же группах подсчитать ***групповой индекс*** (см. раздел «Логические и аналитические индексы») по второму признаку. По значениям этого индекса получить второй ранжированный ряд.

4. Подсчитать необходимый для вашего случая коэффициент ранговой корреляции. Обосновать, почему выбран именно такой, а не другой коэффициент. Проанализировать полученное значение коэффициента.

Выводы из главы 3

1. Начало начал анализа данных — это процессе планирования исследования, этап разработки программы исследования, разработки концептуальной схемы исследования.

2. В процессе построения модели изучения свойства социального объекта продумывается логика поиска простых эмпирических закономерностей. В целом «язык» анализа данных в предполагаемом исследовании определяется только после осмысления логики интерпретации эмпирических закономерностей, т. е. ответа на вопрос: Что и как будем делать, если получим то-то и то-то?

3. Независимо от выбора стратегии анализа (восходящей или нисходящей) социологу необходимы умения первичного анализа, первичной обработки данных. Одномерные распределения, таблицы сопряженности только просты по виду. Социолог может использовать множество «языков» анализа данных при работе с ними.

4. Меры центральной тенденции различны для разных типов шкал. Средняя арифметическая без дисперсии, медиана без квартального размаха, мода без коэффициента качественной вариации для социолога не имеют содержательного смысла.

5. В зависимости от того, с какими из относительных частот работает социолог, он решает разные типы содержательных задач.

6. Изучение связи между признаками (эмпирическими индикаторами или производными от них показателями) - одна из целей анализа. Связь, взаимосвязь трактуются, понимаются по-разному. Потому так много мер (коэффициентов) связи.

7. В таблице сопряженности находится вся информация о взаимосвязи двух признаков.

8. Изучение взаимосвязей невозможно без понимания таких пар понятий: «функциональная - корреляционная связь», «локальные меры связи - глобальные», «сильная связь - слабая», «ложное значение коэффициента - истинное», «направленная связь - ненаправленная», «статистическая зависимость - независимость» и т. д.

9. Меры связи различаются для различных типов шкал и для разного понимания связи.

10. Коэффициенты парной связи целесообразно использовать только в сравнительном контексте в рамках одного и того же исследования. Эффективными являются две стратегии их использования: поиск факторной структуры совокупности признаков; поиск признаков, детерминирующих целевой признак.