Количественные методы в социологических исследованиях

Паниотто Владимир Ильич, Максименко В.С.

Киев , 2003

Монография посвящена описанию логики мер статистического анализа социологической информации, выводу и детальному рассмотрению коэффициентов и статистических показателей, использующихся в социологии. Рассмотрены вопросы обработки социологической информации на программируемых микрокалькуляторах и приведены программы расчета большинства изложенных в книге показателей. Монография содержит наиболее полную сводку статистических таблиц.

Содержание:

ВВЕДЕНИЕ

В нашей стране все более широкое распространение получают конкретные социологические исследования, которые являются источником разносторонней социальной информации, необходимой для успешного решения важных социально-экономических задач, для научного управления общественными процессами.

В последние десятилетия в этих исследованиях интенсивно используются математические методы. Это закономерный этап в развитии социологии. Сегодня социологи уже не сомневаются в том, что в социальных исследованиях необходимо сочетать количественный и качественный анализ, что социология должна применять современные математико-статистические методы так же, как естествознание и экономика.

Однако практическое использование этих методов наталкивается на известные трудности. Как справедливо отмечается в редакционной статье журнала «Коммунист», «...в работе социологов до сих пор недостаточно эффективно используются количественные, математические методы и современная вычислительная техника»[1]. Объясняется это во многом тем, что социологи, как правило, не обладают специальными математическими знаниями, а обслуживающие их математики — знанием предмета исследования.

Для того чтобы найти общий язык с математиком, социолог прежде всего должен понимать смысл, особенности и возможности статистических методов. Однако положение его довольно затруднительно: чисто математические руководства оказываются практически недоступными из-за отсутствия соответствующей подготовки, а руководства,

[3]

[1] Социологические исследования: результаты, проблемы и задачи.— Коммунист, 1980, № 13, с. 82.

разработанные, скажем, для инженеров или биологов, могут быть использованы лишь в ограниченной степени из-за специфики социологического материала.

Очень полезное начинание осуществил в 1968 г. Институт философии АН СССР, издав «Методику и технику статистической обработки первичной социологической информации»[2]. Эта книга, рассчитанная на лиц, не имеющих специальной математической подготовки, была встречена с интересом и несомненно сыграла позитивную роль в самообразовании социологов. Однако она не лишена ряда естественных для первого издания недостатков, которые во многом были устранены авторами при работе над «Рабочей книгой социолога», содержащей разделы, посвященные статистике, и книгой «Статистические методы анализа информации в социологических исследованиях»[3] (авторы последней в большинстве своем принимали участие и в создании первых двух работ).

«Статистические методы анализа...» представляют собой наиболее полное руководство по использованию статистических методов, включающее основные современные методы анализа информации и достаточно широко отражающее диапазон статистических методов, используемых социологом. Вместе с тем изложение здесь (как и в двух предыдущих книгах) носит преимущественно характер готовых рецептов, что делает эту книгу более удобной в качестве справочника для опытного исследователя, чем в качестве пособия для изучения сущности статистических методов.

Солидаризуясь с мнением известного венгерского ученого А. Реньи, мы полагаем, что, изучив только рецепты, нельзя их использовать правильно, а подлинное освоение материала и, следовательно, успешное его применение невозможно без упорного умственного труда[4]. Вот почему мы старались написать книгу, в которой основные меры и формулы *выводятся, подробно анализируются* и *обосновываются.* В нашей книге, как правило, детально рассматриваются условия применимости статистических мер, а также вопросы проверки их значимости. В ряде случаев мы приходим к показателям сначала из качественных (или полукачественных) соображений, а затем даем, по возможности, строгий вывод.

[4]

[2] Методика и техника статистической обработки первичной социологической информации. М., 1968.

[3] Рабочая книга социолога. М., 1977; Статистические методы анализа информации в социологических исследованиях. М., 1979.

[4] Реньи А. Трилогия о математике. М., 1980, с. 94.

Это позволяет читателю, испытывающему затруднения на втором этапе, ограничиться первым и получить тем не менее определенное представление о соответствующем статистическом показателе. Поскольку при выводе и анализе формул не применяется сложный математический аппарат, авторы надеются, что чтение книги не вызовет особых трудностей у широкого круга исследователей.

Уделяя значительное внимание вопросам измерения социальных признаков, мы предлагаем, как нам представляется, детально обоснованную классификацию всех основных статистических мер по уровням фактического измерения. Это должно помочь читателю войти в круг идей статистического анализа и правильно применять соответствующие меры. При таком подходе мы были вынуждены сосредоточить внимание лишь на фундаментальных вопросах, изучение которых поможет социологу разобраться подробнее в специальной литературе и в том материале, который изложен в данной книге конспективно.

В книге содержатся многочисленные примеры, почерпнутые из практики отечественных и зарубежных исследований, а также оригинальных исследований с участием авторов. Эти примеры помогают понять логику применения статистики в социологии, приемы и способы анализа информации, разобраться (что очень важно) в вопросах интерпретации полученных результатов. Читателю предлагается также выполнить ряд упражнений (там, где необходимо, они снабжены указаниями — подсказками, ответами, анализом результатов). Подобная работа поможет уяснить смысл излагаемого материала. Естественно, это требует активного чтения, известных усилий. Изучив эту книгу, социолог не станет математиком (такую задачу и ставить нецелесообразно), однако сумеет, мы надеемся, понять и прочувствовать сущность статистических методов, следовательно, правильно выбрать те, которые нужны для решения возникающих перед ним проблем, грамотно поставить задачу математику и верно проинтерпретировать результаты.

Количественные методы, конечно, не заменят качественный, содержательный анализ, но могут сделать его эффективным. Для того чтобы статистические методы «вели к углубленному пониманию изучаемых явлений, исследователь, их применяющий, должен сам стоять на высоте задачи. Он должен не только владеть инструментом, но также владеть материалом и предметом своего исследования. Он должен быть способен применять технику статистической

[5]

работы к преследуемым целям и имеющимся возможностям. Шаблонное же, механическое использование готовых рецептов, хотя даже и опирающееся на самые точные формулы и самые тонкие математические соображения, ведет не к умножению наших знаний ценой больших, но оправданных затрат труда, а к бесплодному расточению сил и нагромождению числового материала, мало продвигающему вперед наше понимание изучаемых явлений»[5].

Ограниченность объема книги обусловила конспективность некоторых глав (V—VII). Не имея возможности детально излагать весь материал, мы все же сочли необходимым рассмотреть вопросы обработки социологической информации на ЭВМ, проверки статистических гипотез, надежности данных, получаемых социологом, так как в ряде публикаций встречаются ошибки, вызванные недостаточно корректным использованием статистических методов — от планирования выборки до расчета значимости полученных показателей. Например, некоторые авторы склонны придавать значение даже незначительным различиям в полученных данных и трактовать их, не прибегая к тщательной проверке значимости. Проверка значимости представляется нам обязательной для исследователя (в некоторых примерах, основанных на социологических публикациях, мы показали, что определенные неточности допускают иногда даже высококвалифицированные социологи).

Впервые в отечественной литературе рассмотрены вопросы обработки социологической информации на программируемых микрокалькуляторах и приведены программы расчета большинства изложенных в книге показателей. Опыт работы отдела социологических исследований Института философии АН УССР показал высокую эффективность сочетания ЭВМ (для первичной) и программируемых микрокалькуляторов (для большинства видов вторичной обработки информации). Авторы полагают нерациональной ориентацию на преимущественную обработку информации на ЭВМ и выделяют широкий класс задач, для решения которых целесообразней использовать программируемые калькуляторы. Это дает значительную экономию времени (не говоря уже о финансах). Кроме того, работа с калькулятором не требует посредников (программистов, операторов), образуя своеобразную диалоговую систему, позволяющую наилуч-

[6]

[5] Чупров А. А. Основные проблемы теории корреляции. М., 1925, с. 125.

шим образом организовать итеративный процесс анализа информации: «гипотеза — расчет показателей для ее проверки — интерпретация и выдвижение новой гипотезы и т.п.». Направление это является весьма перспективным, так как прогресс в области микроэлектроники предполагает разработку новых типов программируемых калькуляторов и микро-ЭВМ, «равномерно» заполняющих разрыв между обычными калькуляторами и большими ЭВМ. Новая вычислительная техника будет эффективней, чем большие ЭВМ, для подавляющего большинства видов вторичной обработки информации.

Авторы надеются, что книга представит интерес также и для специалистов по использованию статистических методов в социологии. В ней предлагаются некоторые оригинальные приемы анализа информации: оптимизация размещения большого числа полигонов на одном графике, существенно расширяющая традиционные представления о возможностях конденсации информации в графической форме; разработка алгоритмов расчета некоторых статистических коэффициентов для типичных в социологии форм представления первичной социологической информации; нормировка модульного Δ-коэффициента, позволяющего корректно использовать этот показатель для описания связей и др. Монография содержит наиболее полную из опубликованных в нашей литературе сводку статистических таблиц, часть из которых — оригинальна (рассчитана на микрокалькуляторе по составленным авторами программам).

Думается, что книга может быть полезна не только социологам, но и специалистам, изучающим вопросы экономики, психологии, биологии, истории, демографии и др., которые интересуются применением количественных методов в конкретных исследованиях.

В книге принята сквозная нумерация примеров, таблиц и упражнений. Формулы нумеруются внутри каждого параграфа отдельно. Так, (III, 1,2) означает вторую формулу 1-го параграфа 3-й главы.

[7]

Глава I

ИЗМЕРЕНИЕ И АНАЛИЗ РАСПРЕДЕЛЕНИЙ

1. Об измерении в социологии. Классификация социальных признаков по уровням измерения

Количественный анализ применяется при изучении разнообразных форм движения материи, но необходимым условием его эффективности всегда является предварительный качественный, содержательный анализ изучаемых явлений. Как отмечал Гегель, «качество есть непосредственная определенность и с него следует начинать»[6]. Именно качественный анализ определяет постановку задачи, вычленяет предмет исследования, выбирает способы и средства исследования, в частности адекватные задаче количественные методы, использование которых углубляет, делает более конкретным наше знание.

Количественные методы могут быть применены в исследовании лишь после того, как эмпирические данные переведены на язык чисел. Предпосылкой и началом применения количественных методов в социологических исследованиях является измерение. Обычно под измерением понимается «познавательный процесс, в котором определяется отношение одной (измеряемой) величины к другой однородной величине» принимаемой за единицу измерения»[7]. Однако это определение пригодно лишь для измерения количественных[8] (например, стажа, заработной платы и т.п.), а не качественных признаков (например, удовлетворенности, оценки, ориентации и т.п.), так как здесь нет общепризнанных

[8]

[6] Гегель Г. В. Ф. Соч., т. 5, М., 1937, с. 65.

[7] Философская энциклопедия, т. 2. М., 1967, с. 244.

[8] Количественным называется признак, значениями которого служат числа, допускающие сложение; в противном случае признак называется качественным. (Суппес П., Зинес Дж. Основы теории намерений.— В кн.: Психологические измерения. М., 1967, с. 25). После введения понятий уровней измерения различие качественных и количественных признаков станет более ясным.

эталонов и единиц измерения. Поэтому имеет смысл расширить понятие измерения, понимая под ним процедуру приписывания чисел значениям признака. Цель измерения — получить числовую модель, исследование которой могло бы заменить исследование самого объекта. Это возможно лишь тогда, когда свойства модели соответствуют свойствам объекта, т.е. отношения между числами, образующими числовую модель, соответствуют отношениям между изучаемыми свойствами объекта.

Итак, мы понимаем под *измерением особую процедуру, в результате которой возникает числовая модель объекта* (точнее, изучаемых свойств объекта). При измерении, таким образом, устанавливается соответствие между свойствами объекта и свойствами сопоставленных им чисел. Набор свойств объекта и сопоставляемых им чисел называют шкалой[9] (свойства объекта трактуются здесь очень широко, в частности, под набором свойств понимаются также и различные степени интенсивности одного свойства).

В естественных науках предполагается, что всегда можно пользоваться всеми свойствами чисел. Это обстоятельство настолько привычно, скажем, для физики, что пользуются им обычно автоматически; при этом получаются вполне корректные следствия.

Аксиомы арифметики поэтому так оправданы в физическом мире, что создавались в результате отражения, пусть не всегда осознаваемого (вспомним, например, положение И. Канта об априорности математического знания) свойств и отношений этого мира. Как писал Энгельс, само «понятие числа заимствовано исключительно из внешнего мира, а не возникло в голове из чистого мышления»[10]. Поэтому математические, в частности арифметические, понятия сохраняют следы своего происхождения[11]. Для физика, например, естественно, что масса в 15 кг в 3 раза больше, чем масса в 5 кг, и на 10 кг больше последней. Это кажется столь очевидным, что воспринимается как трюизм. Когда же мы переходим в область психологии или социологии, ситуация значительно усложняется. Здесь исследователь нередко рискует произвести такую арифметическую трактовку своих

[9]

[9] В теории измерений под шкалой понимают однозначное отображение эмпирической системы с отношениями в числовую систему с соответствующими отношениями. *(Суппес П., Зиме Дж,* Основы теории измерений..., с. 19; *Пфанцагль И.* Теория измерения. М., 1976, с. 23).

[10] *Маркс К., Энгельс Ф.* Соч., т. 20, с. 37.

[11] *Реньи А.* Трилогия о математике. М., 1980, с. 44.

измерений, которая оказалась бы лишенной всякого смысла[12].

Вот почему со всей определенностью нужно подчеркнуть важность изучения базовых эмпирических отношений, которые в конечном счете определяют допустимые операции с числами, приписанными объектам в каждом конкретном случае. Поясним это примером. Предположим, что мы изучаем удовлетворенность работников своей работой (точнее предприятием, на котором они работают).

Обычно в таких случаях вначале выдвигается содержательная модель данной социальной переменной, скажем, из следующих 5 пунктов:

a) вполне удовлетворен работой;

b) скорее удовлетворен, чем не удовлетворен;

c) промежуточная позиция;

d) скорее не удовлетворен, чем удовлетворен;

e) совершенно не удовлетворен.

В качестве эмпирических референтов соотнесения индивидов с позициями модели могут, например, использоваться ответы на вопросы социологической анкеты. Возможные варианты ответов упорядочиваются по схеме так называемого логического квадрата[13]. Рассмотрим построение шкалы с помощью двух вопросов.

Первый — о переходе на другое предприятие и второй — о возврате (в прожективной ситуации: «Допустим, что Вы некоторое время не работали на заводе. Вернулись бы Вы на него?») имеют варианты ответов: «да», «нет», «не знаю».

Схема «логического квадрата» в нашем случае принимает такой вид:

Варианты ответа на вопрос о переходе	Варианты ответа на вопрос о возврате		
	«Да»	«Не знаю»	«Нет»
«Нет»	*a*	*b*	*f*
«Не знаю»	*b*	*c*	*d*
«Да»	*f*	*d*	*e*

Здесь *a, b, c, d, e,* обозначают соответствующие пункты шкалы, *f* — противоречивые ответы.

[10]

[12] *Решлен М.* Измерение в психологии.— В кн.: Экспериментальная психология. М., 1968, с. 197.

[13] Рабочая книга социолога. М., 1976, с. 232.

Шкалы могут строиться и на большем числе вопросов. Пунктам шкалы и, следовательно, попадающим туда индивидам, приписываются числа *X,* например: 5, 4, 3, 2, 1. Но можно ли считать, что различие в степени удовлетворенности между работниками, попадающими в позиции «*a*» и «*b*», такое же, как между индивидами, попадающими в «*b*» и «*c*», «*c*» и «*d*»? Можно ли утверждать, что индивиды, попадающие в позицию «*b*», вдвое больше удовлетворены, чем те, которые попадают в позицию «*d*»? Ясно, что ответы на эти вопросы должны быть отрицательными. Мы не имеем права пользоваться свойствами равенства интервалов и отношений, так как данные свойства не обеспечены соответствующими свойствами объектов: между ними установлено лишь отношение порядка.

В принципе можно приписать позициям числа X' = 2, 1, 0, —1, —2 (что означает применение преобразования $X \rightarrow X'=X—3$). Числа можно возвести в квадрат ($X \rightarrow X'=X^2$) и вообще: любое монотонное преобразование, не изменяющее последовательности чисел, является в данном случае допустимым. Это обстоятельство необходимо учитывать при выборе статистических мер, осуществлении арифметических операций над числами. И так в каждом конкретном случае.

Приписывание чисел пунктам шкалы, как правило, неоднозначно, т.е. числа допускают определенные группы преобразований, не меняющих их (чисел) свойств.

Тип шкалы можно определить допустимыми группами преобразований ее чисел[14] или допустимыми арифметическими операциями над этими числами[15]. При обоих подходах тип шкалы, или уровень измерения, фактически детерминируется эмпирическими свойствами изучаемой системы.

Теоретически существует бесконечное число типов шкал. Но обычно, когда шкалы различают по уровню измерений — от самых «слабых» к самым «сильным», то выделяют 4 уровня. (4 типа шкал): номинальные (ординарные), порядковые (ординальные), интервальные и, наконец, шкалы отношений (релятивные, или пропорциональные).

Такая классификация, как мы увидим, является одновременно классификацией и по допустимым арифметическим операциям, и по допустимым группам преобразований чисел.

[11]

[14] *Stevens S. S.* On the theory of scales of measurement.— Science, 1946, v. 103.

[15] *Coombs C. H.* Theory and methods of social measurement.— In: *Festlnger L., KatzD.* Research methods In behavioral sciences. N. Y., 1953.

Чем выше уровень шкалы, тем уже круг допустимых преобразований чисел, тем больше арифметических свойств реализуется и, тем самым, шире применяемый статистический аппарат. Для шкал данного уровня можно использовать статистические меры шкал всех предшествующих уровней, но не наоборот.

Познакомимся в общих чертах с основными типами шкал (после изучения статистических мер мы вернемся к шкалам, рассмотрев принципиальный вопрос классификации мер по уровням измерения признаков).

Номинальные шкалы

Для построения этой шкалы необходимо уметь устанавливать отношение равенства (и неравенства) объектов — в смысле рассматриваемого признака — для распределения изучаемой общности на непересекающиеся, дизъюнктивные классы, каждый из которых является отдельным пунктом шкалы. Исследователь должен найти такие эмпирические индикаторы, с помощью которых любой объект можно соотнести с определенным классом, т.е. позицией на шкале. Иногда эта задача решается просто (или сравнительно просто) — установление принадлежности к нации, полу, вероисповеданию и т.д., но зачастую она оказывается далеко не элементарной. Так, длительные поиски предшествовали выделению О.И. Шкаратаном[16] структурных групп, представляющих пункты номинальной шкалы, по которым распределяются члены такой социальной общности, как современное промышленное предприятие. Напомним эти группы:

I — организаторы производственных коллективов;

II — работники высококвалифицированного научно-технического труда;

III — работники квалифицированного умственного труда;

IV — организаторы первичных производственных коллективов;

V — работники высококвалифицированного труда, сочетающие умственные и физические функции при обслуживании сложной техники;

[12]

[16] *Шкаратан О. И.* Социальная структура советского рабочего класса.— Вопросы философии, 1967, № 1; *Шкаратан 0. И.* Проблемы социальной структуры рабочего класса СССР, историко-социологическое исследование. М., 1970; *Шкаратан О. И., Рукавишников В. О.* Социальные слои в классовой структуре социалистического общества.— Социологические исследования, 1977, № 2.

VI — работники квалифицированного физического ручного труда;

VII — работники квалифицированного, преимущественно физического труда, занятые на машинах и механизмах;

VIII — работники нефизического труда средней квалификации;

IX — работники неквалифицированного физического труда.

В расположении структурных групп интуитивно угадывается известный порядок, но интуиция, «угадывающая» по-рядок, не доказывает его наличия. При детальном рассмотрении мы видим, что «нисходящее» расположение групп не всегда оправдывается; так и творческий характер труда, и престиж, и заработная плата, например, работников V и VI групп могут быть выше, чем у работников I или IV (можно привести и другие примеры несоответствия этому порядку). Следовательно, шкала структурных групп остается неупорядоченной, фактически она номинальная.

Другой пример построения номинальной шкалы — выяснение причин текучести работников. Здесь увеличение числа классов (пунктов), желательное в принципе для более детального изучения проблемы, нередко приводит к увеличению ошибок, уменьшению надежности получаемых результатов за счет нарушения требования дизъюнктивности, т.е. приводит к появлению пересекающихся классов. Например, в одной из работ по текучести выделяется, в частности, такая причина увольнения — «решил перейти к друзьям»[17]. Очевидно, что причиной перехода здесь могут быть и условия труда, и жилищно-бытовые условия («там, говорят, скорее квартиру получить можно») и т.д. Другой источник возможных ошибок — использование слов, допускающих очень широкое толкование, например, «семейные обстоятельства» и др.

Обычно рассматриваемые классы укрупняются в блоки, *содержательно* непересекающиеся. При исследовании текучести, выделяются, например, такие блоки: 1) неудовлетворенность условиями трудовой деятельности; 2) неудовлетворенность заработком; 3) неудовлетворенность жилищно-бытовыми условиями,

При этом итоговые данные оказываются ненадежными, так как закладываются ошибки при распределении недизъ-

[13]

[17] Социальные проблемы труда и производства. Москва, Варшава, 1969, с. 229.

юнктивных (пересекающихся) классов в непересекающиеся блоки (ошибки первой стадии классификации).

Отметим, что для обоснованного построения не «очевидной» шкалы представляется перспективным применение методов таксономии[18].

Итак, хотя номинальная шкала обеспечивает только самый слабый тип измерения, процедура ее построения зачастую не тривиальна. Единственное требование, предъявляемое к числам, приписываемым различным классам в случае номинальных шкал — быть *различными.* Очевидно, эти числа могут быть подвергнуты любому взаимно-однозначному преобразованию, то есть от чисел X всегда можно перейти к $X'=f(X)$, где $f(X)$ — закон взаимно однозначного сопоставления. В дальнейшем мы будем для краткости обозначать это так: $X \rightarrow X'=f(X)$. Здесь числа играют роль символов, «ярлыков», их вполне можно заменить, например, любыми буквами, или какими-либо другими знаками. И то, что обычно выбирают для нумерации позиций натуральные числа 1, 2, 3, ... диктуется лишь соображениями удобства, привычки.

Порядковые шкалы

Для построения такой шкалы необходимо уметь устанавливать не только отношения равенства между объектами (по данному признаку), но и отношения *последовательности* — порядка. Это отношения типа «больше, чем», «лучше, чем» и т.д. Далее, как мы видели, выдвигается содержательная модель признака (см., например, шкалу удовлетворенности работой). Эмпирическим референтом могут быть специальный тест (например, набор проективных ситуаций), вопрос (или, чаще, система вопросов) социологической анкеты, и т.д. С помощью референтов объекты социальной общности соотносятся с пунктами шкалы. Каждому пункту может быть приписано некоторое число. Между этими числами имеют место те же отношения, что и между объектами. Ясно, что и в случае порядковых шкал приписывание чисел неоднозначно.

Этими числами могут быть и 1, 2, 3, 4, ... и 1, 4,9, 16, ... и 1, 3, 5, 7 ... и т.д., т.е. любое преобразование $X \rightarrow X'=\varphi(X)$, где $\varphi(X)$ — монотонно возрастающая функция,

[14]

[18] См. главу VI.

которая не изменит свойств чисел, приписанных пунктам (свойствам объекта). Известна лишь их последовательность, но не расстояния между ними. Вообще говоря, расстояния между пунктами шкалы не равны (подчеркиваем, что использование рангов может породить иллюзию равенства!), мы не только не можем сказать, *во сколько раз* одно значение признака больше другого, но и на сколько. Следовательно, и числа фактически не несут такой информации.

Понять это помогает простой пример. Рассмотрим такую порядковую шкалу, как итоговое распределение мест в турнирной таблице спортивных состязаний. Ясно, что в общем случае расстояния между этими позициями разные (например, первый «оторвался» от второго больше, чем второй от третьего и т.д.). Конечно, судьи и болельщики знают расстояния (в очках) между различными позициями. В случае порядковой шкалы мы находимся в положении человека, который знает только распределение мест и не может узнать количество очков, набранных разными участниками.

Отметим, что ранги определяют относительную интенсивность качества, но не «абсолютную» величину ее. Ценность шкал этого типа в том, что они устанавливают порядок, а недостаток в том, что этот порядок не является метрическим.

Приведем несколько примеров. Порядковой является шкала ветров Бофорта. Ее пункты: «штиль», «легкий ветер», «свежий», «крепкий», «шторм», «ураган». Каждый из них имеет качественное определение (эмпирический референт). Эти определения основаны на действиях, производимых ветром. Порядок расположения пунктов шкалы фиксируется числом баллов. Так, «легкий ветер», например, 3 балла, «крепкий» — 7, «шторм» — 10 баллов. Сами эти числа фиксируют не абсолютную интенсивность свойства (силы ветра), а лишь отношения последовательности между пунктами. Их нельзя, например, складывать, но можно сравнивать (больше — меньше).

В минералогии существует эталонная шкала твердости из 10 пунктов, каждому из которых приписывается число — от 1 до 10. Пункты расположены в порядке возрастания твердости (шкалируемый признак). Единица соответствует тальку, 10 — алмазу. На этой шкале любому минералу отводится место с помощью такой процедуры: данный минерал располагается между тем, который он царапает, и тем, который царапает его. Так возникает порядковая шкала.

[15]

Педагогическая система балльных оценок — пример порядковой шкалы: мы не можем сказать, что знания студента, получившего 5, на столько больше знаний студента, получившего 4, на сколько знания последнего больше знаний получившего 3. Нельзя также, например, сказать, что знания получившего 4 вдвое больше знаний получившего 2 (очевидна также размытость позиций этой шкалы), хотя можно в идеале утверждать, что знания получившего 5 больше знаний получившего 4 и т.д. Это же относится ко всем балльным шкалам. Поэтому: *шкалы, построенные с помощью балльных оценок, строго можно рассматривать лишь как порядковые, но не метрические.* Число случаев, когда это предается забвению, достаточно велико. Между тем, практически все современные шкалы в социологии и психологии — номинальные и порядковые.

Интервальные шкалы

В основе построения интервальной шкалы лежит эмпирическая процедура, позволяющая определить равенство *дистанций* между *парами* объектов (разумеется, наряду с определением равенства и порядка объектов). Если эта процедура найдена, числа, приписываемые пунктам шкалы, обладают таким свойством: равенство интервалов чисел отвечает равенству эмпирических интервалов, т.е. интервалов между интенсивностями свойств у рассматриваемых пар объектов. Поэтому свойства чисел, приписанных объектам, не изменяются при линейном преобразовании $X \rightarrow X' = aX + b$. Действительно, если для двух пар объектов *A, B* и *C, D* (так мы условно обозначим эти объекты), $X_B - X_A = X_D - X_C$, то и $X'_B - X'_B = X'_D - X'_C$. Но при этом, если $\frac{X_B}{X_A} = \frac{X_D}{X_C}$, то отсюда не следует, что $\frac{X'_B}{X'_A} = \frac{X'_D}{X'_C}$, т.е. нет равенства отношений.

В преобразовании $X \rightarrow X' = aX + b$ есть два неопределенных параметра — *a* и *b,* и поэтому можно сказать, что в шкале интервалов произвольны начало отсчета *(b)* и единица измерения *(a).*

Интервальными являются, например, все температурные (Цельсия, Реомюра, Фаренгейта) шкалы, кроме абсолютной (Кельвина). Как известно, температура по Фаренгейту связана с температурой по Цельсию соотношением $X' = 32 + 1{,}8X$. Выбирая разные значения *X*, можно легко

[16]

убедиться, что в этой шкале нет равенства отношений. У температурных шкал произволен выбор точки отсчета — нуля (в шкале Цельсия, совершенно условно, это температура замерзания воды, например), произволен и масштаб (цена деления разная у шкал Цельсия, Фаренгейта и Реомюра).

Интервальными являются также календарные шкалы. Даты одного и того же события в разных календарях тоже связаны между собой линейным законом.

Подобные шкалы в социологии редки, ими пользуются для измерения пространственных и временных положений объектов. Зато нередки псевдоинтервальные шкалы (шкала Терстоуна, «термометр» общественного мнения и т.д.), т.е. шкалы, по некоторым признакам напоминающие интервальные, но по сути являющиеся порядковыми.

Шкалы отношений

Базовая эмпирическая процедура построения такой шкалы заключается в установлении равенства отношений между *парами* объектов по изучаемому признаку (разумеется, наряду с отношениями равенства, порядка, равенства интервалов между парами объектов). Числа, приписываемые объектам в этом случае, обладают свойствами равенства отношений, т.е. практически удовлетворяют всем арифметическим аксиомам. Допустимые преобразования чисел теперь суть преобразования подобия: $X \rightarrow X' = aX$ $(a > 0)$, т.е. фиксировано начало отсчета, можно лишь менять масштаб, единицу измерения. Следовательно, приписав определенное число какому-нибудь объекту, тем самым фиксируем числа, приписываемые всем другим аналогичным объектам. Классическим примером такой шкалы являются абсолютная (кельвиновская) температурная шкала, а также обычная числовая шкала счета. Если $a=1$, то шкалу называют абсолютной. В качестве примера таковой приводят обычно шкалу счета (если считать единицами, а не десятками, сотнями и т.д.).

В социологии такие шкалы используются для измерения «физических» величин — времени (стаж, возраст), счета (заработная плата, доход, премия), когда «экспериментально» определен нуль — начало отсчета. Пример абсолютной шкалы — социометрический статус члена группы (число полученных им выборов).

В зависимости от типа шкалы применяются те или иные методы статистического анализа, после ознакомления с

[17]

которыми мы вернемся к классификации статистических мер по выделенным уровням социологического измерения. Отметим, что различие интервальных шкал и шкал отношений для социологических исследований практически несущественно, эти два типа шкал часто объединяют в один тип и называют *метрическими* шкалами (метр от греческого *μετρου* — мера). Особенностью метрических шкал является наличие единицы измерения и допустимость операции сложения. Возвращаясь к определению количественных и качественных признаков, можно сказать, что количественными называются признаки, измеренные с помощью метрических шкал, а качественными — с помощью шкал более низкого уровня (в частности, номинальных и порядковых). Это определение подчеркивает относительность различий качественных и количественных признаков и связь этих различий с уровнем измерения (можно, например, считать, что до изобретения термометра температура была качественным признаком, так как измерялась с помощью порядковой шкалы: горячий, теплый, комнатный, прохладный, холодный, ледяной).

Конкретные шкалы не всегда легко отнести к тому или иному типу. Например, некоторые авторы считают образование (в годах обучения) количественным признаком. Но при строгом подходе в силу разнокачественности одного года обучения в школе, в техникуме и в вузе, этот признак нужно рассматривать как измеренный в порядковой шкале (это следует иметь в виду при выборе статистических мер). То же самое касается квалификации рабочих, измеряемой разрядами. С другой стороны, эти шкалы так же, как, например, балльные оценки знаний в школе, содержат все же больше информации, чем чисто порядковые: между пунктами шкалы существует некоторое, хотя и приближенное равенство. Ведь преподаватель, выставляющий балл, старается использовать шкалу как метрическую, поэтому, например, изменение системы баллов с 2, 3, 4, 5 на 2, 3, 20, 21 рассматривалось бы как некорректное увеличение расстояния между удовлетворительными и хорошими знаниями. Такие шкалы находятся, следовательно, где-то между метрическими и порядковыми (их иногда называют псевдоинтервальными или псевдометрическими), поэтому при строгом подходе корректно применение лишь статистики для порядковых шкал, но в некоторых случаях возможно (при известной осторожности) использование статистики для метрических шкал.

[18]

2. Табулирование. Вариационные ряды. Графики. Приемы наглядного представления социологических данных

Предположим, что мы опросили некоторое множество респондентов с помощью следующей анкеты[19]:

Социологическая анкета

1. Укажите, пожалуйста, Ваш пол:

мужской 1

женский 2

2. Удовлетворены ли Вы своей профессией?

полностью удовлетворен 1

скорее удовлетворен, чем нет 2

затрудняюсь ответить 3

скорее не удовлетворен, чем удовлетворен 4

неудовлетворен 5

3. Укажите, пожалуйста, доход на одного члена Вашей семьи – руб.

Здесь представлены три типа признаков (1-й вопрос порождает номинальную шкалу, 2-й — порядковую и 3-й — метрическую), поэтому на этом примере можно рассмотреть основные специфические для социологии методы представления данных[20]. Прежде всего, сведем информацию к обозримому виду, перенеся данные из анкет в специальную таблицу 1.

Такого рода таблицы называются *матрицами данных.* Дальнейшие преобразования информации направлены на то, чтобы сделать ее более наглядной, представить в более компактной форме. С этой целью подсчитывают, сколько индивидов обладают данным значением признака. Значение признака называют *вариантом,* а число лиц, обладающих данным значением,— его *частотой.* Варианты вместе с частотами образуют *вариационный ряд* данного признака, или *распределение* по данному признаку (в табл. 2 представлены вариационные ряды признаков «пол» и «удовлетворенность профессией», или распределение опрошенных по признакам «пол» и «удовлетворенность профессией»).

[19]

[19] Предлагаемая анкета носит иллюстративный характер, по существу, это фрагмент реальной социологической анкеты, содержащей обычно десятки (или даже сотни) вопросов, в том числе: контактных, функционально-психологических, контрольных и т.п. (См., например, *Ядов В. А.* Социологическое исследование. М., 1972; *Ноэль Э.* Массовые опросы. М., 1978; и др.).

[20] Исключение представляют данные, порождаемые социометрическими вопросами, — методы их анализа рассмотрены в гл. VI.

Таблица 1

Условный пример: данные опроса 284 респондентов

Номера индивида (анкеты)	Признак		
	Пол	Удовлетворенность профессией	Доход
1	1	2	80,0
2	2	1	75,3
3	2	5	65,4
.	.	.	…
283	2	4	82,3
284	1	2	95,0

Таблица 2

Распределение опрошенных по признакам «пол» и «удовлетворенность профессией», частости и проценты

Показатель	Признак							Всего
	Пол		Удовлетворенность профессией					
	Номер варианта ответа							
	1	2	1	2	3	4	5	
Частота	104	180	81	83	19	61	40	284
Относительные частоты, или доли, частости	0,37	0,63	0,28	0,29	0,07	0,21	0,14	1
Процент	37	63	28	29	7	21	14	100

Наряду с вариационными рядами в табл. 2 содержатся также частости и проценты. *Частостями* называют частоты, разделенные на сумму частот по данному признаку, другое название — *относительные частоты,* или *доли частот* (в данном примере сумма частот для признаков «пол» и «удовлетворенность профессией» равна 284), *проценты* представляют собой умноженные на сто частости (доли).

Представить компактно данные, полученные по метрическим шкалам, таким способом, как правило, не удается из-за большого количества вариантов, поэтому для построения распределения диапазон изменения признака разбивают на интервалы и подсчитывают, сколько индивидов имеют значение признака, лежащее в границах каждого интервала (табл. 3).

[20]

Из таблицы ясно, что 2 индивида имеют доход до 65 руб., 32 — от 65 до 74 руб. и т.д. Отметим, однако, что использованные нами значения округлены до целых, т.е. значение 74,3 руб., например, отнесено к интервалу 65—74, а значение 74,6 к интервалу 75—84. Условимся цифру 5 округлять до высшего разряда, т.е. 74,5 до 75 и, следовательно, относить 74,5 к интервалу 75—84. В некоторых работах используются интервалы с совпадающими границами, т.е. в данном примере это были бы границы: до 65, 65—75,

Таблица 3

Распределение опрошенных по признаку «доход»

Показатель	Номер интервала						Всего
	1	2	3	4	5	6	
	Граница интервала, руб.						
	до 65	65-74	75-84	85-94	95-104	105 и выше	
Частота	2	32	50	181	11	8	284

75—85, 85—95, 95—105, 105 и выше. В этом случае 74,5; 74,8; 74,9, например, относятся к интервалу 65 — 75, а значения 75,0; 75,1 и т.д.— к интервалу 75 — 85. (Это важное замечание будет учтено при выводе формул для вычисления медианы и квантилей).

Для описания вариационных рядов введем следующие обозначения. Значения признака *X y* отдельных индивидов, т.е. варианты, обозначим через x_i , i= 1, 2, ..., *N,* где *N* — общее число индивидов, или объем совокупности. (Для краткости в дальнейшем мы будем писать $i = \overline{1, N}$). Некоторые варианты могут повториться: например, на предприятии имеется ряд работников с образованием 10 классов и т.д. Пусть различных вариантов *k (k<N),* а обозначение x_i при $i=\overline{1, k}$ соответствует теперь различным вариантам. Общее число индивидов с $X = x_i$ мы будем обозначать $N(X_i)$ или просто N_i (пока рассматривается один признак это возможно). Ясно, что $\sum_{i=1}^{k} N(x_i) = \sum_{i=1}^{k} N_i = N$ .. Величина N_i является частотой, а $v_i=N_i/N$ – частота варианта x_i. Очевидно, что $\sum_{i=1}^{k} v_i = 1$.

[21]

Варианты вместе с частотами образуют вариационный ряд (одномерное распределение признака), который может быть дискретным (в случае номинальных и порядковых признаков, а также для некоторых метрических, например, «число детей в семье», «разряд» для рабочих и т.п.) или непрерывным (для метрических признаков). В случае, если варианты расположены в порядке убывания или возрастания, вариационный ряд называется упорядоченным (ранжированным). Как правило, непрерывные признаки указанным способом преобразуют в дискретные путем введения интервалов. Величина интервала называется интервальной разностью.

Если обозначить левую границу некоторого *l*-го интервала через x'_l, а правую — через x''_l, то ширина интервала, или интервальная разность, равна $I_l = x''_l - x'_l$. Эта формула верна лишь в случае, если границы соседних интервалов совпадают, т.е. $x''_l = x'_{l+1}$. Когда границы интервалов не совпадают (как в табл. 3), то $I_l = x''_l - x'_l + 1$. Например, ширина 3-го интервала равна не 84—75=9, а 10, так как в интервал попадают, как указывалось выше, значения от 74,5 до 84,5 (84,5—74,5=84—75+1=10). Величину $\frac{1}{2}(x''_l + x'_l)$ (для интервалов с совпадающими границами) или $\frac{1}{2}(x''_l + x'_l + 1)$ (для интервалов с несовпадающими границами) назовем серединой или центром интервала. Для нашего примера середина интервала равна $\frac{1}{2}(84+75+1)=80$.

Основным приемом представления и анализа социологических данных является построение одномерных (вариационные ряды) и двумерных распределений признаков (реже 3-мерных и *n*-мерных распределений) или, другими словами, распределений опрошенных по одному, двум, трем и более признакам.

Одномерное распределение

А. **Классификационные и качественные признаки** (номинальные и порядковые шкалы). Допустим, что нам известно одномерное распределение *N* респондентов по некоторому признаку *X,* имеющему *k* градаций (вариантов):

Вариант	x_1	x_2	…	x_k
Частота	N_1	N_2	…	N_k

[22]

Чаще всего одномерное распределение изображается с помощью полигонов и гистограмм распределения. На оси абсцисс откладываются k точек, на оси ординат — значения N_i; соединив их ломаной линией, получим *полигон* распределения, если же построить столбики высотой N_i — получим *гистограмму*. Полигоны и гистограммы можно строить не только с использованием частот, но и частостей и процентов.

Таблица 4

Распределение населения СССР по уровню образования в 1979 г. (см. Население СССР. М., 1980 г.)

Уровень образования	Абсолютные показатели, тыс. чел.	Процент
Неполное среднее	52488	37,7
Среднее общее	45099	32,4
Среднее специальное	23439	16,9
Высшее незаконченное	3235	2,3
Высшее оконченное	14826	10,7
Всего	139089	100

Рассмотрим на примере, как строятся указанные виды графиков.

Пример 1. Построим полигон и гистограмму распределения для данных, приведенных в табл. 4.

На рис. 1, отражающем эти данные, изображены две оси ординат — на одной из них отложены абсолютные величины, на другой — проценты. Форма графиков не зависит от вида показателя (частоты, частости или проценты), откладываемого на оси ординат. Полигон (по соглашению) изображают как замкнутую кривую.

Б. **Количественные признаки** (интервальные шкалы и шкалы отношений). Принципиальных различий в построении одномерных распределений количественных признаков по сравнению с изображением качественных признаков нет, но есть некоторые особенности, связанные с тем, что для количественных признаков приобретает смысл понятие ширины интервала. Прежде чем перейти к обсуждению этого вопроса, введем некоторые определения.

Частоту, приходящуюся на единицу интервала (для l-го

[23]

интервала $\rho_l = \frac{N_l}{I_l}$, назовем *плотностью распределении,* а частость, приходящуюся на единицу интервала — *относительной плотностью распределения.* Особо важную роль играет это понятие в случае неравных интервалов, на чем мы в дальнейшем специально остановимся.

Нам также понадобится понятие *накопленной,* или *кумулятивной, частоты* (частости). Накопленная частота по-

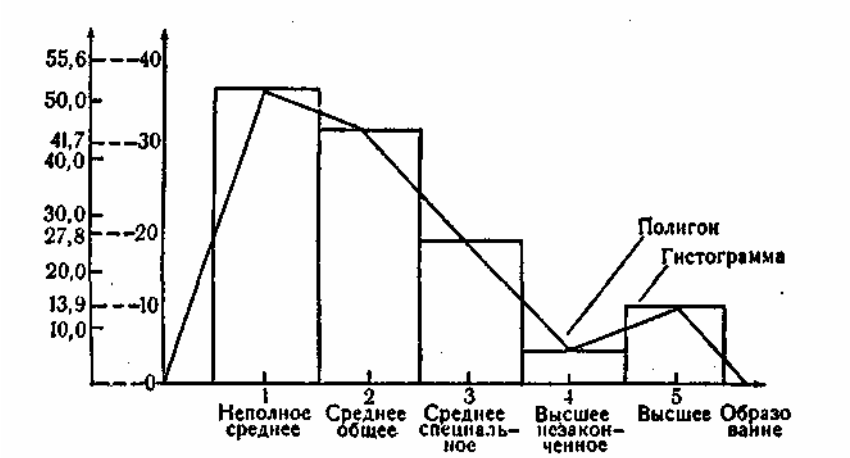


Рис. 1. Уровень образования населения СССР (1979 г.)

казывает число индивидов, у которых варианты не больше (меньше либо равны) данного значения признака.

Скажем, для *l*-ого интервала накопленная частота $F_l = \sum_{i=1}^{l} N_i$ — показывает, у какого числа индивидов $X \le x_l$ или, другими словами: сколько всего индивидов с $X=x_1$, $X=x_2$, ..., $X=x_l$. Очевидно $F_k=N$. Кумулятивная частость $f_l = \sum_{i=1}^{l} v_i (l \le k)$ и соответственно $f_k=1$. Тогда ρ_l в процентах равна $\frac{v_l}{I_l}$. 100%.

В конкретных исследованиях нередко используются неравные интервалы. Так как велик диапазон возможных значений, например, возраста работников (свыше 50 лет), то при равных интервалах в случае разумного числа пунктов (10—12) будет слишком большой интервальная разность (около 5 лет), это не позволит достаточно точно изучить по-

[24]

ведение работников разного возраста, особенно молодых (в старших возрастных группах, как показывают исследования, влияние возрастных различий на поведение несколько ниже) Увеличение же дробности, желательное для детального изучения, приводит к очень большому числу пунктов (25—30), существенно затрудняющему анализ материала. Выходом из этого положения является компромиссный вариант: малые интервалы выбираются для групп молодых

Таблица 5

Распределение по возрасту работников Одесского судоремонтного завода им. 50-летия Советской Украины (1971 г.)

Граница интервала, лет	Середина интервала, x_i	v_i, %	f_l, %	I_l	ρ_l, %
16–17	16,5	2,4	2,4	2	1,20
18–19	18,5	5,8	8,2	2	2,90
20–21	20,5	5,1	13,3	2	2,55
22–24	23,0	10,9	24,2	3	3,63
25–30	27,5	15,3	39,5	6	2,55
31–40	35,5	30,2	69,7	10	3,02
41–50	45,5	18,3	88,0	10	1,83
51–60	55,5	8,5	96,5	10	0,85
Свыше 60	65,5	3,5	100,0	10	0,35

работников, а большие — для работников старших возрастных групп.

В настоящее время в социологической литературе обсуждается проблема стандартизации основных измерительных процедур. Дел в том, что данные, получаемые разными исследователями, зачастую несопоставимы (или крайне ограниченно сопоставимы). В значительной мере это результат отсутствия соглашений между исследователями по поводу измерения различных признаков. Практически получается, что число разных градаций одного и того же признака не намного меньше числа исследователей. Осознавая эти трудности, экспертная служба ИСИ АН СССР провела опросы социологов страны, в частности по проблеме «Возраст в конкретных исследованиях». Анализ результатов позволяет дать некоторые рациональные рекомендации для социологов-практиков[21].

[25]

[21] *Петренко Е.С., Ярошенко Т.М.* Социально-демографические показатели в социологических исследованиях. М., 1979, с. 40—49.

Обратимся к примеру, иллюстрирующему данные выше определения.

Пример 2. В таблице 5 приведено распределение по возрасту работников Одесского судоремонтного завода им. 50-летия Советской Украины (1971 г.). Как видим, при построении распределения использовались неравные интервалы. Рассмотрим, например, интервал 20—21, сюда мы относим индивидов, возраст которых от 19,5 до 21,5, т.е. ширина

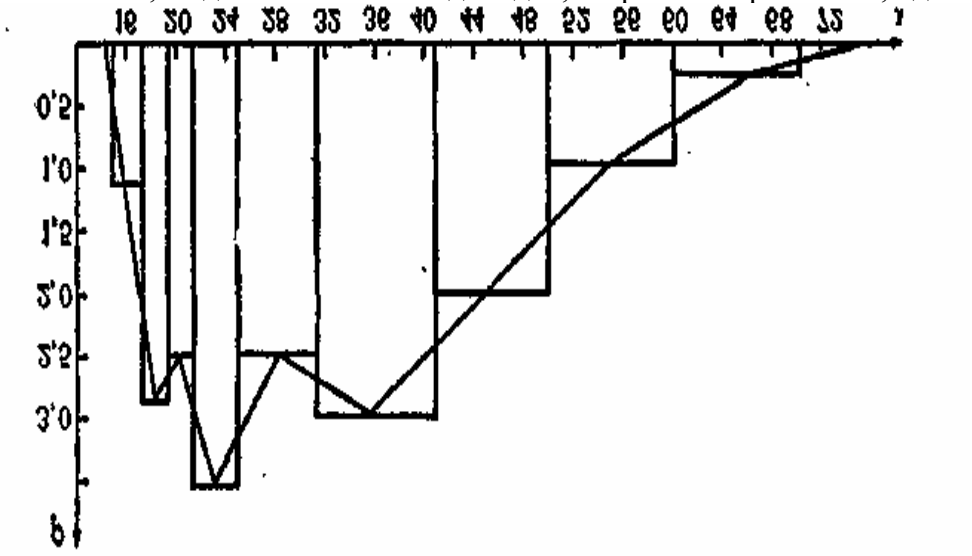


Рис. 2. Полигон и гистограмма распределения при неравных интервалах

интервала 2 года, в интервал 25—30 попадают индивиды, возраст которых от 24,5 до 30,5, т.е. ширина его 6 лет.

Если правая граница предыдущего интервала совпадает с левой последующего (например, в случае интервалов 16— 18, 18—20, 20—22 и т.д.), то следует указать, к какому из них относить граничное значение (в данной книге мы относим его к верхнему интервалу). Отметим, что возникающие трудности, если такое указание не сделано, зачастую преувеличиваются: вероятность того, что мы опрашиваем индивида в день его рождения порядка тысячных долей ($^1/_{365}$)[22].

Из-за наличия неравных интервалов, для построения полигона распределения данных, приведенных в таблице 5, по оси ординат откладывают уже не N_i (или v_i), а плотности ρ_i. Аналогично строится и гистограмма (рис. 2). Отметим, что площадь каждого прямоугольника равна $I_l\rho_l=N_l$, а сумма площадей всех прямоугольников равна N.

Плотность изображается на гистограмме так, как если бы

[26]

[22] О понятии вероятности см. Приложение 1.

она была постоянной внутри интервала. Обычно этого нет, ρ_l — это средняя плотность на интервале. Ясно, что чем меньше интервал, тем ближе полигон к фактическому изменению плотности распределения в зависимости от изменения признака. Для непрерывных признаков в пределе, когда $I_l \to 0$, мы получили бы плавную кривую изменения плотности распределения, которую называют теоретической кривой распределения. Очевидно, площадь, ограниченная кривой распределения, равна 1, если на оси ординат откладывать частости. В дальнейшем мы подробнее остановимся на кривых распределения.

Еще один графический способ изображения вариационного ряда — кумулятивная кривая (ее называют также кумулятой, или кривой накопленных частот). Кумулята строится аналогично полигону, но координаты точек теперь (x_l, F_l) либо (x_l, f_l) т.е. абсциссы те же, а ординаты — накопленные, или кумулятивные, частоты. Ясно, что кумулята — неубывающая кривая.

Упражнение 1. Построить кумуляту по данным табл. № 5.

Кривая, построенная по точкам с координатами (F_l, x_l), называется огивой Гальтона[23].

Упражнение 2. Для нашего примера построить огиву.

Форма статистического распределения (вариационного ряда) — вид его графика. Например, полигона. Проанализируем полигон рис. 2. Вначале с увеличением возраста увеличивается плотность распределения. Затем — провал, он связан с уходом молодежи в армию (на обследуемом предприятии работают в основном мужчины). Затем плотность снова возрастает: на предприятие приходят отслужившие. Второй провал связан с историческими условиями жизни страны — эхо войны, следствие низкой рождаемости и выживаемости детей в военные годы (это станет ясно, если сопоставить соответствующие *x c* годом опроса, с течением времени этот провал, естественно, сдвигается вправо. Затем плотность распределения монотонно убывает с увеличением возраста, что естественно.

Полигон — ломаная кривая. Вид полигона зависит от числа различных вариантов. Предел, к которому стремится полигон при увеличении числа вариантов, плавная кривая, которая может быть описана с помощью некоторого аналити-

[27]

[23] Кумулята и огива позволяют быстро определить долю лиц, обладающих более высоким (или низким) значением, чем любое фиксированное значение признака. Например, медиана является ординатой такой точки огивы, абсцисса которой равна 0,5.

ческого выражения: $y=y(x)$. Разные распределения описываются с помощью различных функций.

Познакомимся с некоторыми часто встречающимися формами распределений.

Распределение может описываться монотонной — убывающей или возрастающей — функцией типа изображенных на рис. 3 (а и б соответственно).

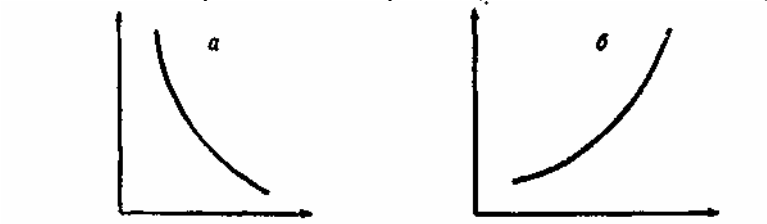


Рис. 3. Монотонно убывающая (а) и монетонно возрастающая (б) функции

Рис. 4. *U*-образное (а) и *J*-образное (б) распределение

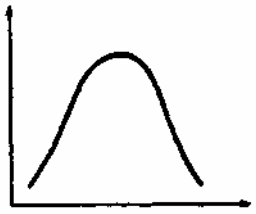


Рис. 5. Колоколообразное распределение

Примером здесь может служить распределение работников по стажу работы на данном предприятии: чем боль-ше стаж, тем меньше работников (это связано с трудовыми перемещениями, с текучестью: уходом «старых» и приходом новых работников).

Распределение может быть *U*-образным (частный случай —*J*-образным, см. рис. 4а и 46 соответственно): например, распределение по удовлетворенности трудовой деятельностью (как правило, часто оказывается меньше всего работников, занимающих на шкале удовлетворенности промежуточную позицию).

Своего рода обратным *U*-образному является так называемое колоколообразное распределение (рис. 5), встречающееся довольно часто в конкретных исследованиях: например, распределение людей по росту, весу, по заработной плате («крайности» встречаются редко). Если частоты вариантов, симметричных относительно центрального, при-

[28]

мерно одинаковы, то распределение называется симметричным, в противном случае — асимметричным. На рис. 6 (а—г) показаны примеры асимметричных распределений:

Одновершинные распределения называются унимодальными, двувершинные — бимодальными и т.д. Многовершинные распределения встречаются реже одновершинных.

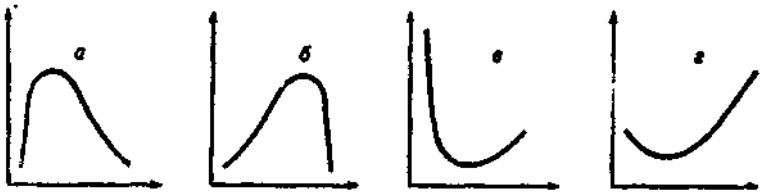


Рис. 6. Асимметричные распределения

Часто встречаются колоколообразные распределения, хотя и не всегда в «чистом» виде: эмпирическое распределение может быть близким к колоколообразному. Особо важную роль в статистике играет распределение, получившее название *нормального* (§ 3 этой главы).

Двухмерные распределения (комбинационные таблицы)

Рассмотрим следующую таблицу, представляющую собой двухмерное распределение[24] по признакам «тип рабочего места» и «удовлетворенность зарплатой» данных выборочного почтового опроса жителей Киева (табл. 6). Такого рода таблицы иногда называют комбинационными, так как в них отражена информация о комбинации двух (в данном случае) или большего числа признаков.

На пересечении *i*-й строки и *j*-го столбца этой таблицы стоит число респондентов, имеющих *i*-е значение первого и одновременно *j*-е значение второго признака, а также процент, который составляет это число от суммы элементов строки. Фактически, таблица представляет собой 4 вариационных ряда (если не считать итогового распределения, которое приведено для удобства пользования таблицей). Поэтому данные этой таблицы можно изобразить на одном графике в виде 4-х полигонов, используя для каждого свой цвет или вид линии (сплошная, пунктирная и т.п.).

[29]

[24] Формализованное описание двухмерных распределений мы приведем ниже при рассмотрении корреляционной таблицы (гл. II, § 1, табл. 15).

Упражнение 3. Начертить график, представляющий данные двухмерного распределения признаков, приведенные в таблице 6.

Таблица 6

Двухмерное распределение данных почтового опроса жителей г. Киева, абсолютная величина и процент

Признак «тип рабочего места»			Признак «удовлетворенность работой»					Всего
Труд	Квалификация	Номер варианта	Удовлетворен	Скорее да, чем нет	Трудно ответить	Скорее нет, чем да	Неудовлетворен	
			Номер варианта					
			1	2	3	4	5	
Физический	Низкая	1	86 32.3	23 8,6	27 10,2	31 11,6	99 37,2	266 100
	Средняя и высокая	2	393 41,2	110 11,5	117 12,3	114 11,9	221 23,1	955 100
Умственный	Не требующая высшего образования	3	182 23,9	64 8,4	89 11,7	121 15,9	306 40,2	762 100
	Требующая высшего образования	4	245 25,8	123 12,9	106 11,1	168 17,7	309 32,5	951 100
	Всего		906 30,9	320 10,9	339 11,6	434 14,8	935 31,9	2934 100

Если один из признаков двухмерного распределения количественный, мы имеем возможность для каждого значения качественного признака рассчитать средние арифметические[25] и таким образом «сжать» информацию, как бы свести ее к одномерному распределению (например, если второй признак не «удовлетворенность», а «доход», то можно было бы рассчитать средний доход для каждого из четырех типов

[30]

[25] Средние арифметические рассматриваются в следующем параграфе.

Таблица 7

Удовлетворенность респондентов различными сторонами своей работы

Удовлетворенность	Рабочее место				Все группы	Ранг	Среднее квадратическое отклонение	Ранг
	Физического труда		Умственного труда					
	Низкой квалификации	Высокой и средней квалификации	Не требующего высшего образования	Требующего высшего образования				
1	2	3	4	5	6	7	8	9
1. Содержанием труда	0,46	0,61	0,57	0,64	0,60	3	0,079	3
2. Режимом труда	0,54	0,46	0,49	0,55	0,50	4	0,042	6
3. Размером оплаты	-0,06	0,18	-0,20	-0,09	-0,03	7	0,160	1
4. Возможностями повышения квалификации	0,26	0,35	0,22	0,20	0,26	6	0,066	5
5. Отношениями с коллегами	0,92	0,90	0,87	0,84	0,87	1	0,035	7
6. Отношениями с руководителями	0,81	0,73	0,74	0,64	0,72	2	0,070	4
7. Удаленностью работы от места жительства	0,60	0,45	0,54	0,35	0,46	5	0,109	2
8. Возможностями улучшения жилищных условий	-0,17	-0,21	-0,21	-0,14	-0,18	8	0,034	8

рабочих мест). На графике в этом случае будет лишь один полигон распределения: на оси абсцисс — качественный признак, по оси ординат откладываются средние значения количественного признака.

Часто так поступают не только для количественных, но и для качественных признаков, измеренных с помощью порядковых шкал: пунктам шкалы приписываются определен-

ные баллы и находится средний балл[26], или индекс (подробнее этот вопрос будет рассмотрен в § 3). Так, приписав удовлетворенным балл l, тем, кто скорее удовлетворен, чем нет — 0,5, затрудняющимся ответить — 0, тем, кто скорее неудовлетворен, чем удовлетворен — (—0,5) и, наконец, неудовлетворенным — балл (—1), получим для каждого типа рабочих мест следующие средние баллы (индексы) удовлетворенности:

Тип рабочего места	1	2	3	4
Индекс удовлетворенности зарплатой	- 0,06	0 ,18	- 0,20	- 0,09

Таким образом, данные «сжались» до одной строки и могут быть изображены в виде одного полигона (по оси абсцисс — типы рабочих мест, по оси ординат — индексы удовлетворенности). На одном графике можно изобразить данные целого ряда таблиц двухмерных распределений. Так, в проведенном нами опросе работающего населения г. Киева была получена информация об удовлетворенности респондентов различными сторонами работы, или элементами рабочей ситуации (содержанием и режимом труда, зарплатой и т.п.). Индексы удовлетворенности по восьми двухмерным распределениям респондентов для признаков «тип рабочего места» и «удовлетворенность элементом рабочей ситуации» (одно из них было приведено в табл. 6), сведены в таблицу 7. Но прежде, чем перейти к построению графика, сформулируем некоторые общие принципы изображения нескольких полигонов на одном рисунке. Целесообразно рассмотреть отдельно два случая: а) изображаются два полигона; б) три и более.

В первом случае исследователь ставит перед собой цель наглядно представить различия между двумя группами респондентов (или какими-либо двумя другими объектами). При этом на оси абсцисс откладываются значения признака, а полигоны представляют собой распределения каждой из групп по этому признаку или значения некоторого показателя. Значения признаков на оси абсцисс целесообразно откладывать упорядоченными по убыванию разности ординат полигонов. Поясним сказанное примером. В исследова-

[26] Как будет показано в гл. III, при строгом подходе эта операция не совсем корректна, так как опирается на некоторые непроверенные предположения. Тем не менее практика применения индексов в социо-логии показывает, что для приближенных оценок их использование час-то правомерно.

нии межличностных оценок, проведенном В. Шубкиным, Ю. Карповым и Г. Кочетовым[27], каждому из индивидов предлагалось оценить всех членов своего коллектива (в том числе и себя) по семи группам качеств:

I – интеллектуальные качества (одаренность, глубина знаний и т.п.);

II – деловые качества (умение привлечь людей и т.п.);

III – импульсно-волевые свойства (сдержанность, эмоциональность и т.п.);

IV – моральные качества (доброта, скромность и т.п.);

V – качества, характеризующие мотивы поведения (альтруизм, стремление к истине и т.д.);

VI – качества, характеризующие отношения к жизни (оптимизм, юмор и т.п.);

VII – качества, характеризующие физическую привлекательность

По каждой из групп качеств были найдены коллективная оценка (т.е. оценка данного человека другими) и самооценка (т.е. средняя самооценка членов коллектива). Полученные данные представлены на рисунке, заимствованном из книги В. Шубкина[28] (рис. 7, а). Он дает определенное представление о полученных в результате исследования данных (видно, например, что самооценка выше всего по моральным качествам и качествам, характеризующим отношение к жизни, т.е. п. IV. и VI, что различия оценки и самооценки выше по п. IV, чем по п. III и V и т.д.). Но многие различия оценок и самооценок на графике «не читаются». Например, неясно, по каким пунктам больше различия — по I или по VII, по IV или по VI и т.п.

Чтобы сделать график наглядней и информативней, мы вычли для каждого пункта из коллективной оценки самооценку и расположили качества личности на оси абсцисс по убыванию этой разности (см. рис. 7, б). Интерпретация такого графика существенно облегчается: слева расположены качества личности, для которых оценка других выше, чем самооценка (это интеллектуальные качества и качества, характеризующие физическое совершенство, т.е. п. I и VII, причем по первому из них различия больше), а справа те, по которым индивид оценивает себя выше, чем коллектив (п. IV и VI, а также п. V, III и II, по которым различия приблизительно равны между собой и существенно ниже, чем различия по п. IV и V). Отметим, что при таком способе

[33]

[27] *Шубкин В. Н.* Социологические опыты. М., 1970, с. 110—151. [м] Там же, с. 127.

[28] Там же, с.127.

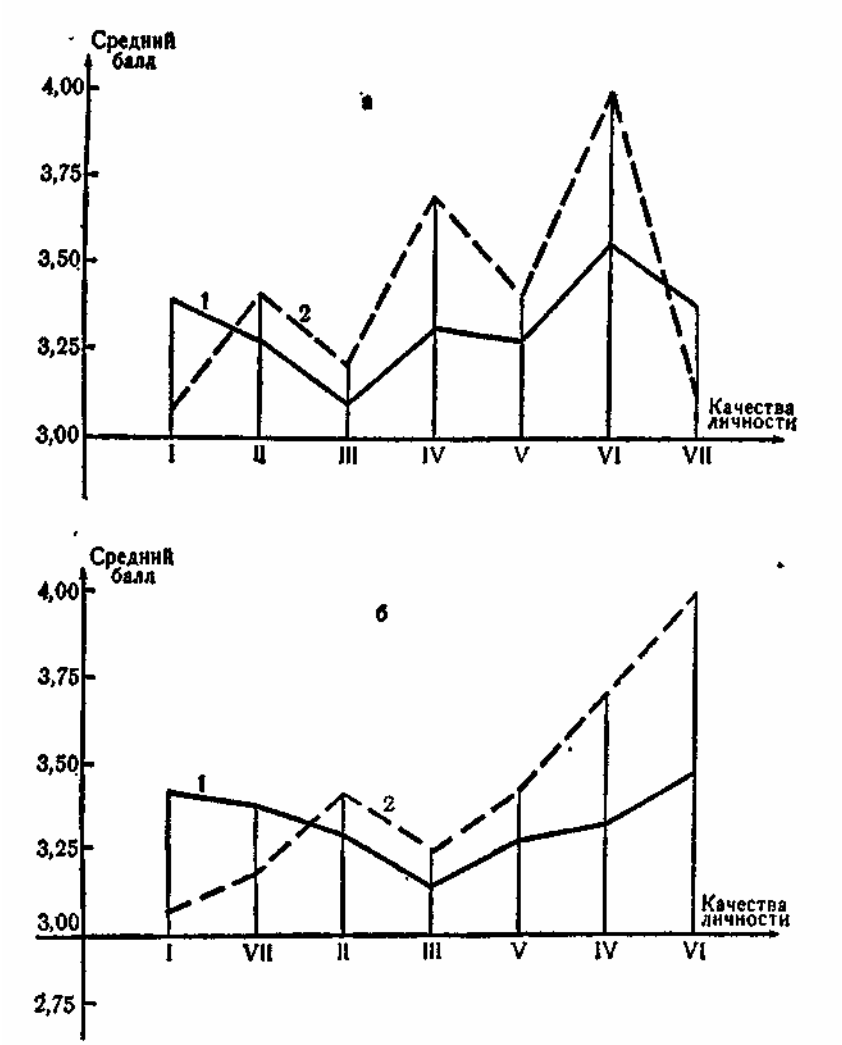


Рис. 7. Коллективная оценка и самооценка качеств личности: график «а» построен неудачно, график «б» — удачно.
Обозначения: *1* — коллективная оценка, *2* — самооценка

построения даже самые запутанные графики с большим количеством пересечений приобретают достаточно простой вид: они содержат *не более одного пересечения,* причем до пересечения один показатель выше другого, а после пересечения наоборот.

[34]

Рассмотрим второй случай — изображение трех и более полигонов на одном графике. Теперь повышение информативности в зависимости от целей анализа осуществляется двумя путями. Первый из них — когда нас интересует преж-

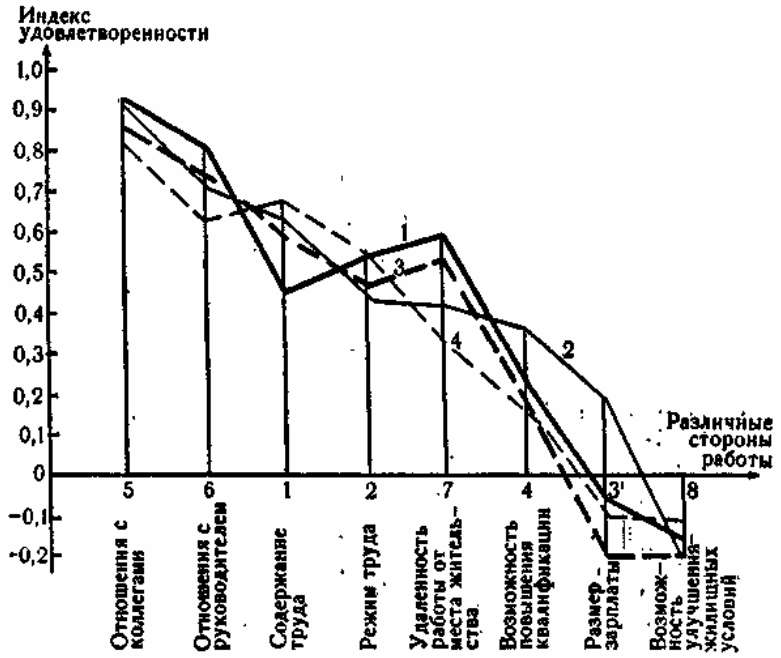


Рис. 8. Удовлетворенность респондентов различными сторонами своей работы (работающее население г. Киева, 1979.). Пункты оси абсцисс расположены по убыванию степени удовлетворенности всех опрошенных.

Обозначения: *1* — лица физического труда низкой квалификации, *2* — лица физического труда средней и высокой квалификации, *3* — лица умственного труда, не требующего высшего образования, *4* — лица умственного труда, требующего высшего образования.

де всего значения изучаемых показателей, а затем уже различия показателей у разных групп респондентов — заключается в расположении пунктов на оси абсцисс по убыванию некоторого усредненного значения изучаемого показателя. Рассмотрим это на примере изображения данных таблицы 7. Предположим, что в первую очередь нас интересует степень удовлетворенности респондентов различными сторонами своей работы, а потом уже различия в удовлетворенности разных групп респондентов. В 6-й колонке таблицы приведены данные об удовлетворенности, рассчитанные для всех

3*

[35]

групп в целом, т.е. для всего массива опрошенных, не расчлененного на группы по характеру труда и уровню квалификации. На рис. 8 на оси абсцисс различные стороны работы представлены в порядке убывания индексов удовлетворенности для всего массива (т.е. в соответствии с рангами,

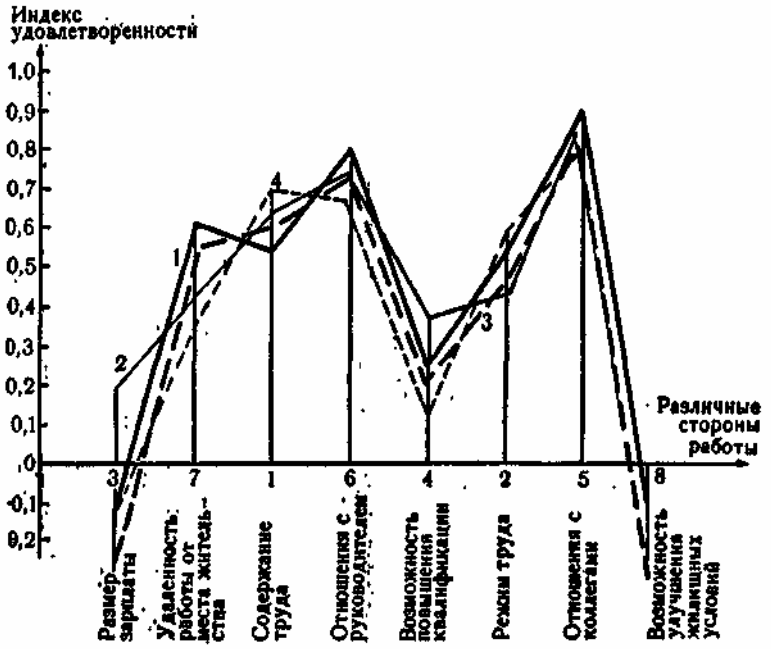


Рис. 9. Удовлетворенность респондентов различными сторонами своей работы (работающее население г. Киева, 1979 г.). Пункты оси абсцисс расположены по убыванию различий между изучаемыми группами.

Обозначения: *1* — лица физического труда низкой квалификации, *2* — лица физического труда средней и высокой квалификации, *3* — лица умственного труда, не требующего высшего образования, *4* — лица умственного труда, требующего высшего образования

приведенными в колонке 7). При этом способе изображения мы имеем возможность при интерпретации обращать внимание прежде всего на наиболее важные, «проблемные» моменты» изучаемых явлений (в данном случае на стороны работы с наиболее низкими показателями удовлетворенности). В тех же случаях, когда нас интересуют прежде всего различия между группами (например, разработка социальных или экономических мероприятий, направленных на уменьшение различий между некоторыми группами респондентов), пункты располагаются по убыванию различий

[36]

между изучаемыми группами. На рис. 9 данные таблицы 7 изображены таким способом. В качестве показателя различий между группами принято среднее квадратическое отклонение[29] (см. колонку 8). На оси абсцисс стороны работы упорядочены по убыванию этого показателя (т.е. в соответствии с рангами колонки 9). В этом случае интерпретация представленных данных проходит иначе, чем в предыдущем. Из рисунка видно, что различие групп по удовлетворенности возможностями улучшения жилищных условий минимально, затем идет удовлетворенность отношениями с коллегами и т.д. Если на рис. 8 было удобно интерпретировать среднюю удовлетворенность и отклонения от нее, то с помощью (рис. 9) удобно интерпретировать различия между группами.

Кроме полигонов и гистограмм, существуют и другие виды графиков, которые используются, однако, значительно реже. В книге Дж.Гласса и Дж.Стенли[30] приводится пример 15-ти различных способов изображения одних и тех же данных. Там же предложены некоторые общие рекомендации для построения графиков[31]. Вместе с тем отметим, что процесс построения графиков плохо формализуется и требует творческого подхода и критического восприятия общих рекомендаций. Нам, в частности, кажется нецелесообразным замыкание полигонов распределения, ухудшающее «чтение» графиков. С другой стороны, неправомерным представляется достаточно распространенное мнение, что на одном графике не следует размещать более трех полигонов, так как целый ряд линий на графике сливается[32]. Сформулированные нами приемы построения графиков вытекают из противоположных соображений. Совпадение полигонов повышает наглядность, облегчает описание сходства и различия: чем больше сливающихся, тем меньше отличающихся точек и тем легче чтение графика (например, из рис. 8 видно, что три группы работников примерно одинаково удовлетворены содержанием труда, у четвертой — работники физического труда низкой квалификации — удовлетворенность этим элементом

[37]

[29] Для этой цели можно использовать также другие меры вариации (см. § 4 этой главы) и коэффициенты корреляции (например, коэффициент Чупрова между признаком «тип рабочего места» и признаками, характеризующими удовлетворенность сторонами работы).

[30] *Гласс Дж., Стэнли Дж.* Статистические методы в педагогике и психологии. М., 1976, с. 42—43.

[31] Там же, с. 54.

[32] Там же, с. 60.

рабочей ситуации значительно ниже; видно также и то, что все группы примерно одинаково оценивают отношения с коллегами, возможность улучшения жилищных условий и т.д.). Думается, что на одном графике вполне можно изображать до 7—8 полигонов распределения.

Завершая изложение способов представления данных, отметим, что построение графиков не только важная часть исследовательской работы, необходимая для повышения наглядности результатов и передачи другим известной автору информации, но и инструмент анализа: продуманный подход к построению графика, стремление сделать его информативным и наглядным позволяют лучше понять структуру полученных данных, глубже проникнуть в сущность изучаемого явления.

3. Меры центральной тенденции

Как мы видели, вариационный ряд может быть описан с помощью набора величин x_i, N_i ($i=\overline{1,k}$). Однако оперирование с полным набором затруднительно. Для удобства изучения необходимо ввести величину, которая, учитывая особенности данного ряда, была бы сводной, итоговой. Такую величину называют *средней.* Средняя не может полностью заменить ряд. Оперируя с нею, мы теряем часть информации, но отражаем типичное для данной совокупности в данных условиях. Средняя характеризует уровень ряда, его центральную тенденцию.

Чтобы средняя величина была действительно обобщающей характеристикой, улавливающей закономерность, она должна применяться к достаточно однородной совокупности. Выведение средних для неоднородной совокупности может привести к бессмысленному результату, например, метко спародированному Г.Успенским усреднению, когда «миллионщик Колотушкин» и «просвирня Кукушкин», имеющий грош, владеют «в среднем по полмиллиону». Такие средние огульны, фиктивны. (Заметим, что в некоторых случаях даже огульная средняя может быть показательной. Например, памятные «четверть лошади» — столько в «среднем» приходилось в царской России на одну ревизскую душу).

Стал классическим пример разоблачения Лениным статистиков народнического толка, выводивших средние для всего крестьянства, не желая видеть, что оно неоднородно, что часть его принадлежит к сельской буржуазии, часть — к

[38]

батракам. Очевидно, «средние», характеризующие крестьянство «в целом», не могли быть научными.

Итак, вычислению средних должно предшествовать обоснованное выделение в изучаемой совокупности достаточно однородных групп.

Говоря о средней, чаще всего имеют в виду среднюю арифметическую $M = \frac{1}{N}\sum_{i=1}^{N}, x_i = \frac{1}{N}\sum_{i=1}^{k} N_i x_i$ (суммируем до N

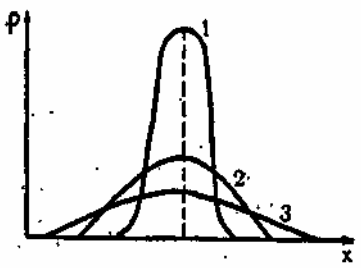


Рис. 10. Распределение с одинаковыми средними

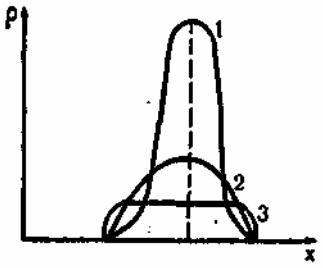


Рис. 11. Распределения с одинаковыми средними и вариационным размахом

или до k). Если все варианты совпадают, то x_i=M, колеблемости (варьирования) нет. Обычно, конечно, $x_i \neq M$. Как же охарактеризовать колеблемость? Простейшей мерой может служить так называемый *вариационный размах* $R=x_{max}–x_{min}$. Для изображенных на рис. 10 распределений такой показатель достаточно эффективен. Все три распределения имеют одинаковые средние $M_1=M_2=M_3$. Ясно, что минимальная колеблемость у распределения 1, максимальная у 3. Как видно из графика, $R_1<R_2<R_3$. Вариационный размах определяется, однако, лишь крайними значениями признака и не отражает колеблемости остальных вариантов. Три распределения, представленных на рис.11, имеют одинаковые R (и M), но явно разные колеблемости. Кроме того, встречаются ситуации, когда вариационный размах в принципе не может быть достаточно достоверно определен (например, доход семей в капиталистических странах — см. пример № 3 этого параграфа). Что же можно еще использовать для описания колеблемости? Величина $X_i—M$ характеризует вклад, вносимый в колеблемость i-ым вариантом.

[39]

Вклад всех вариантов, казалось бы, естественно описать с помощью $\sum_{i=1}^{N}(x_i - M)$. Однако, как легко видеть с учетом определения *M*, эта величина всегда обращается в нуль, следовательно, она не может быть принята в качестве меры колеблемости. Мы получаем нуль из-за взаимной компенсации отклонений разных знаков, т.е. вправо и влево относительно *M.* Наверное, целесообразно освободить отклонения от знаков (в самом деле, ведь и отклонения влево, и отклонения вправо — колеблемость, следовательно, они должны равноправно входить в искомый показатель). В простейшем случае это можно осуществить, переходя к величине $(x_i–M)^2$, которая нивелирует различие «правых» и «левых» отклонений вариантов от *M,* а для полного вклада к $\sum_{i=1}^{N}(x_i - M)^2$. Для сопоставимости различных распределений нужно перейти ко вкладу, приходящемуся на долю одного наблюдения: $\frac{1}{N}\sum_{i=1}^{N}(x_i - M)^2 = D$; эта величина называется *дисперсией,* ее размерность есть квадрат размерности признака. За меру колеблемости естественно принять величину $\sigma = \sqrt{D}$, которая имеет ту же размерность, что и сам признак; она называется *среднеквадратичным (или стандартным) отклонением.* Если колеблемости нет, все $x_i=M$ и $\sigma = 0$. Если а мало, то *M* хорошо представляет ряд, он достаточно однороден. Чем больше о, тем больше колеблемость.

Итак, а показывает, на сколько в среднем отклоняется каждый вариант от *M.* Допустим, что мы сравниваем признаки, имеющие одинаковую размерность. Например, это могут быть общий трудовой стаж, стаж на данном предприятии и т.д. Если одинаковы *M,* то колеблемость больше у того признака, у которого больше о. Если одинаковы о, то это, вообще говоря, не означает, что одинаковы колеблемости. В этом случае колеблемость там меньше, где больше *M.* Для сопоставлений, очевидно, следует перейти к относительному показателю. Таковыми является *коэффициент вариации,* $C_v = \frac{\sigma}{M}100\%$. Сравнивая C_v для общего трудового стажа и стажа на данном предприятии, мы можем сопоставить колеблемость данных признаков индивидов изучаемой общности. Пока речь шла о признаках одинаковой размерности; если же сопоставляемые признаки имеют раз-

личную размерность, то использование коэффициента вариации является единственно возможным способом сравнения колеблемостей. Примерами такого типа являются сопоставления колеблемостей образовательного и квалификационного уровней работников данной профессиональной группы, аналогично для стажа и квалификации, зарплаты и стажа и т.д., в зависимости от стоящей перед исследователем задачи.

Свойства средней арифметической величины.

1. Если все варианты увеличить (или уменьшить) в а раз, то M увеличится (или уменьшится) во столько же раз.

Упражнение 4. Показать самостоятельно (для этого нужно использовать свойства сумм — см. Приложение № 2).

2. Если все варианты увеличить на одно и то же число, то и M увеличится на то же число.

Упражнение 5. Показать самостоятельно. Указание: для этого нужно сделать переход $x_i \rightarrow x'_i = x_i + a$ и вычислить среднее x'_i с использованием свойств сумм, как и в упражнении № 4.

3. Сумма произведений отклонений вариантов от M на частоты равна нулю.

В самом деле, с учетом определения M имеем:

$$\sum_{i=1}^{k} N_i (x_i - M) = \sum_{i=1}^{k} N_i x_i - MN = 0 .$$

4. При уменьшении (или увеличении) частот в одно и то же число раз средняя арифметическая не изменяется.

Упражнение 6. Показать справедливость утверждения самостоятельно.

5. Если совкупность *(N)* разбита на s непересекающихся классов ($N = \sum_{r=1}^{s} N_r$, здесь N_r – число индивидов в r-ом классе), то общая средняя $M = \bar{x}$ равна средней арифметической групповых средних $\overline{x_r}$ ($\overline{x_r} = \frac{1}{N_r} \sum_{i=1}^{k} x_i P_{ri}$, где P_{ri} — число индивидов с $X = x_i$ в r-ом классе), взятых с весами N_r.

В самом деле, по определению,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} x_i N(x_i) = \frac{1}{N} \sum_{r=1}^{s} \sum_{i=1}^{k} x_i P_{ri} = \frac{1}{N} \sum_{r=1}^{s} N_r \overline{x_r} ,$$

что и требовалось показать.

[41]

Упражнение 7. Показать, что

$$\bar{x} = \alpha \frac{1}{N} \sum_{i=1}^{k} \frac{x_i - c}{\alpha} N(x_i) + c \quad \text{(I,3,1)}$$

(Для этого нужно воспользоваться свойствами 1 и 2.)

Упражнение 8. Пусть на заводе три цеха: A, B, и C. Допустим, что средний стаж на данном рабочем месте для работников цеха A — 3,8 года, для работников цеха B — 4,0 года, для работников цеха C — 4,2. Чему равен средний стаж на рабочем месте для всего предприятия в целом? Это зависит от того, сколько работников в каждом из цехов. Пусть в цехе A — 100 человек, в цехе B — 400, в цехе C — 500. Тогда средний стаж для всего предприятия равен:

$$3{,}8\frac{100}{1000} + 4\frac{400}{1000} + 4{,}2\frac{500}{1000} = 4{,}1 \text{ (года).}$$

Такое среднее называется *взвешенным*.

Перейдем к изучению других средних.

Медиана Me — значение признака, которое приходится на центральный (средний) член ранжированного ряда.

У одной половины членов ряда значения признака меньше, чем у среднего, у другого — больше. Допустим, что в отделе главного механика работает 9 человек, возраст которых соответственно: 18, 18, 27, 30, 34, 35, 37, 40, 63 (в годах). Тогда, согласно определению, *Me*=34 года: это возраст работника с условным номером 5. Из оставшихся у половины (№ 1 — 4) возраст меньше, у половины (№6 — 9) больше, чем медианный. Допустим, что в отделе главного бухгалтера 6 человек, возраст которых: 19, 23, 38, 42, 54, 67. По определению принимают, что $Me = \frac{38+42}{2} = 40$. Теперь вообще нет работника с медианным возрастом, ко ровно у половины индивидов возраст меньше, чем *Me,* а у другой — больше.

На медиану влияют лишь центральные, срединные значения признака. Если концы распределений — левый или правый — определены ненадежно, то это не исказит *Me,* поможет исказить *M,* которое зависит от *всех* значений признака.

Заметим, что в некоторых ситуациях применение *M* вообще оказывается невозможным, и *Me* выступает в роли средней, репрезентирующей ряд. Это относится к качественным признакам.

Как вычислить медиану в случае интервального ряда?

[42]

Рассмотрим кумулятивный ряд, т.е. ряд накопленных частот. Медианный интервал — тот, на который приходится 0,5N наблюдений. Пусть его номер l, тогда $N_l=F_l-F_{l-1}$. Все эти варианты заключены между x'_l и x''_l. Мы не знаем точных значений каждого из вариантов, поэтому в простей-шем случае естественно предположить, что внутри интервала все они расположены равномерно, т.е. прирост частоты пропорционален приросту интервала:

$$N_l : I_l = (0{,}5N - F_{l-1}) : (Me - x'_l)$$

Теперь

$$Me = x'_l + I_l \frac{0{,}5N - F_{l-1}}{N_l}. \quad \text{(I,3,2)}$$

Проиллюстрируем это графически:

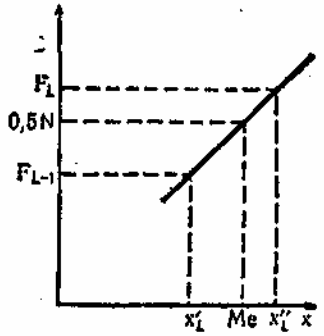


Рис. 12. Определение медианы

Если перейти к частостям, выраженным в %, то:

$$Me = x'_l + I_l \frac{50 - f_{l-1}}{v_l} \quad \text{(I,3,2')}$$

Упражнение 9. По данным примера № 2 вычислить медианный возраст работников.

Мода M_0 — наиболее часто встречающееся в данной совокупности значение признака.

Можно сказать и так: мода — вариант с наибольшей частотой.

Когда продавец говорит о «среднем покупателе», то он, возможно, и не осознавая этого, по существу имеет в виду модального. Мода не отражает степени модальности, сама по себе она не несет информации о том, насколько распространено данное значение признака.

[43]

В отличие от *M* и *Me, Mo* может представлять и классификационные признаки. Можно указать модальную национальность данного государства (например, в СССР это русские), модальную профессию на предприятии или в отрасли и т.д., хотя бессмысленно говорить о средней арифметической или медианной профессии, национальности и т.д.

Как отмечалось, если распределение имеет один максимум, его называют унимодальным (мода — абсцисса макси-

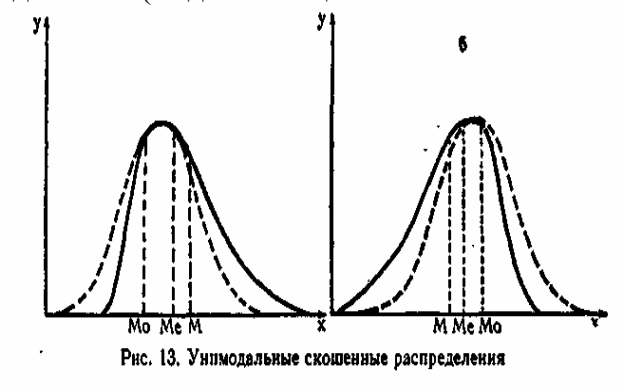


Рис. 13. Унимодальные скошенные распределения

мума), если два, то бимодальным и т.д. Теперь проясняется смысл этих названий. Возрастное распределение населения, например, в отсутствии войн, эпидемий и т.п., обычно имеет колоколообразный вид. У симметричных унимодальных распределений *M=Me—Mo*.

Перейдем к рассмотрению скошенных унимодальных распределений. Сопоставим их с базовым симметричным, которое будем изображать пунктирной кривой. У распределения на рис. 13а «поднят» правый, но «опущен» (по сравнению с симметричным) левый конец. На *Mo* края не влияют (она определяется максимумом, который не изменился, по условию), ее положение не меняется. Как мы видели, на *M* влияют все значения, следовательно, *M* сдвинется, причем в сторону больших значений *X* (поднят правый конец!).

Me тоже сдвигается, но так как она определяется не столько значениями признака, сколько частостей, а в «хвостах» (концах) концентрация события невелика, то и сдвиг *Me* относительно небольшой. Отсюда становится понятным указанное на рисунке взаимное расположение *Mo, Me, M: Mo<Me<M.*

[44]

Упражнение 10. Показать, что если поднят левый конец, то *M<Me< Mo* (рис. 136).

Итак, если *M, Me, Mo* совпадают (либо близки), то распределение симметрично (либо близкое к симметричному). Если же они значительно разнятся и *Mo > Me,* то имеет место левая асимметрия, если *Mo<Me* — правая.

(Замечание: для контроля вычислений можно использовать то, что *Me* всегда между *Mo* к *M).*

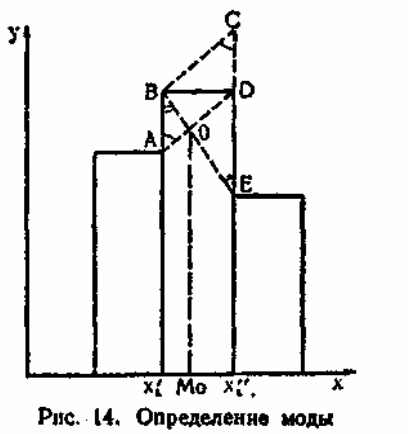


Рис. 14. Определение моды

До сих пор мы предполагали, что умеем вычислять *Mo.* Как же это делать практически в случае наиболее часто встречающихся в социологии интервальных рядов? Прежде всего нужно найти интервал с наибольшим числом наблюдений. Отметим, что при неравных интервалах во избежание ошибок от частоты нужно перейти к плотности. Интервал с наибольшей частотой при равных интервалах (или с наибольшей плотностью при неравных) и есть модальный. Пусть его номер *l.* Естественно предположить, что внутри этого интервала частоты распределены «в согласии» с соседними интервалами: если левый столбик диаграммы (рис. 14) выше, то *Mo* ближе к x_l' если правый, то к x_l''. По определению, в качестве медианы принимается абсцисса точки 0 — пересечения отрезков *BE* и *AD*, удовлетворяющая указанному предположению[33]. Пусть $Mo = x_l' + \Delta x$. Для нахождения Δx проведем (*BC*)II(*AD*) до пересечения с продолжением (*DE*) в точке *C*.

Из подобия треугольников *A0B* и *BCE:*

$$\frac{\Delta x}{I_l} = \frac{|AB|}{|CD| + |DE|};$$

$$|AB| = N_l - N_{l-1}$$

аналогично

$$|CD| + |DE| = |AB| + |DE| = 2N_l - N_{l-1} - N_{l+1}$$

[45]

[33] Отметим, что мы тем самым доопределили *Mo* (!).

Следовательно,

$$Mo = x'_l + I_l \frac{N_l - N_{l-1}}{2N_l - N_{l-1} - N_{l+1}}. \text{ (I,3,3)}$$

Пример 3. Рассмотрим вычисление средних доходов *(M, Me, Mo)* семей США[34] (1959 г.).

Средняя арифметическая $M \approx 6500$ в данном случае мало показательна, ибо здесь усредняются «тигры и кошки»,

Таблица 8

Распределение годового дохода семей США

Годовой доход в долларах	v_i Частость, %	f_i Кумулятивная астость, *%*
до 2000	14	14
от 2000 до 4000	21	35
от 4000 до 6000	23	58
от 6000 до 8000	18	76
от 8000 до 10000	10	86
от 10000 до 15000	9	95
свыше 15 000	5	100

что порождает, пользуясь словами В. И. Ленина, «иллюзию благоденствия».

Для уяснения ситуации вычислим *Me* и *Mo.* Медианный и модальный интервалы у нас совпадают, это 4000–6000:

1) именно на этот интервал приходится максимальная частота (23%);

2) на этот же интервал приходится 50% наблюдений.

Из (I,3,2'): *Me* = 5300, т.е. у 50% семей доход на 20% ниже среднего арифметического. Из (I,3,3) *Mo* = 4600, т.е. наиболее часто встречающийся доход примерно на 30% ниже среднего арифметического.

Упражнение 11. Какой процент американских семей в 1959 г. имел доход ниже, чем средний арифметический?

Ответ: 63%

Упражнение 12. Каков процент семей с доходом ниже модального? Ответ: 42%

Пример 4. По данным табл. 9 о распределении роста 1000 взрослых рабочих-мужчин вычислить *M*, *Mo* и *Me*. Полагая C=165,5, a=3, имеем, используя формулу

[46]

[34] Самуэльсон П. Экономика. М., 1964, с.196

(I,3,1): $M = 165{,}5$. Согласно (I, 3,2): $Me = 164 + 3 \times \frac{500 - 403}{201} = 165{,}5$ (см), а по (I, 3,3): $Mo = 165{,}2$ см. Таким образом, *M, Me* и *Mo* практически совпадают.

Начертим гистограмму (рис. 15). Мы видим, что она, как и следовало ожидать, почти симметрична.

Таблица 9

Вычисление среднего арифметического, моды и медианы

X	x_i	N_i	$x_i - C$	$\frac{x_i - C}{\alpha}$	$\frac{x_i - C}{\alpha} N_i$	$\left(\frac{x_i - C}{\alpha}\right)^2 N_i$	F_i
1	2	3	4	5	6	7	8
143–146	144,5	1	–21	–7	–7	49	1
146–149	147,5	2	–18	–6	–12	72	3
149–152	150,5	8	–15	–5	–40	200	11
152–155	153,5	26	–12	–4	–101	416	37
155–158	156,5	65	–9	–3–	–195	585	102
158–161	159,5	120	–6	–2	–240	480	222
161–164	162,5	181	–3	–1	–181	181	403
164–167	165,5	201	0	0	0	0	604
167–170	168,5	170	3	1	170	170	774
170–173	171,5	120	6	2	240	480	894
173–176	174,5	64	9	3	192	576	958
176–179	177,5	28	12	4	112	448	986
179–182	180,5	10	15	5	50	250	996
182–185	183,5	3	18	6	18	108	999
185–188	186,5	1	21	7	7	49	1000

Рассмотренный пример позволяет перейти к очень важному распределению – нормальному. Сперва несколько вводных замечаний. Рассмотрим последовательность $S_n = \left(1 + \frac{1}{n}\right)^n$. Легко видеть, что $S_1 = 2; S_2 = 2{,}25; S_3 = 2{,}39, \ldots, S_{100} = 2{,}69$.

Упражнение 13. Используя логарифмирование, вычислить S100.

Предел, к которому стремится S_n при неограниченном увеличении n, оказывается некоторым иррациональным числом, которое обозначается через *e*. Можно показать, что, например, с точностью до 4 знаков после запятой e = 2,7183.

[47]

Говорят, что величина *X* распределена нормально, если теоретическая кривая плотности распределения описывается функцией типа

$$y = y_0 e^{-(x-V)^2/2q^2} \quad \text{(I, 3, 4)}$$

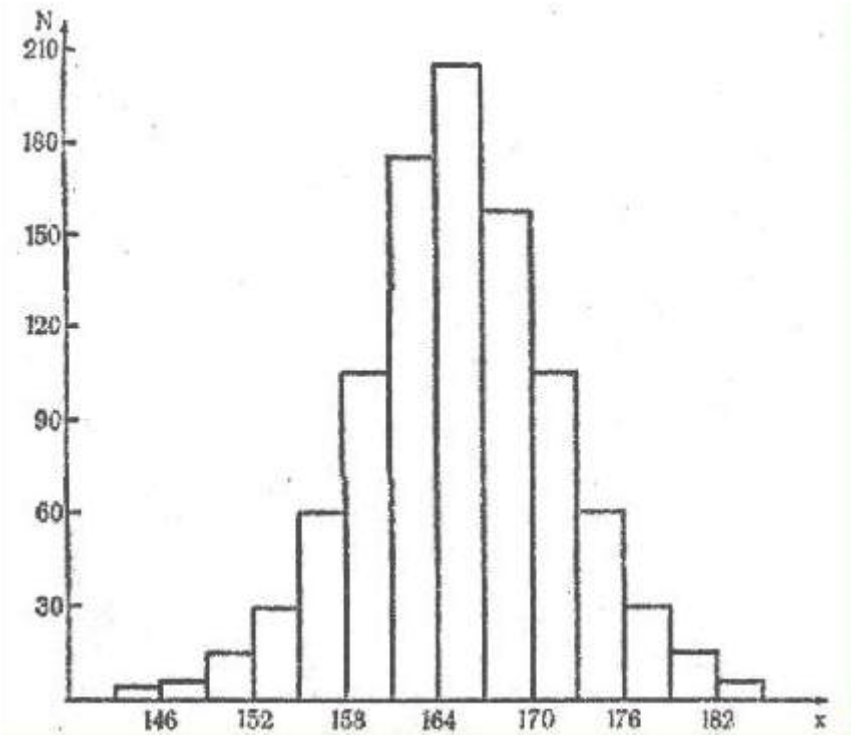


Рис. 15. Гистограмма распределения мужчин по росту

где *M* – среднее арифметическое (абсцисса, относительно которой симметрична кривая), σ – среднее квадратическое, а y_0 – максимальная ордината, равная $\frac{1}{\sigma\sqrt{2\pi}}$. Эта колоколообразная кривая ассимптотически приближается к оси *x*. Нормальное распределение полностью определяется величинами *M* и σ. Вид кривой не зависит от *M,* которое определяет лишь положение максимума, его абсциссу (ордината – y_a). Ясно, что $M = M_0 = M_e$. Форма (вид) кривой определяется величиной σ. Вся площадь, ограниченная этой кривой и осью абсцисс, равна **N**, если по оси ординат отложены частоты, или если – частости (именно из этого условия получено значение $y_0 = \frac{1}{\sigma\sqrt{2\pi}}$), или 100% (если –

[48]

проценты). Оказывается, 68,27 % наблюдений заключено между $M - \sigma$ и $M + \sigma$; 95,45% между $M - 2\sigma$ и $M + 2\sigma$; 99,73% – между M-3σ и M +3σ.

Составлены специальные таблицы, в которых для любого z (взятого с определенным шагом) указано, какая площадь, ограниченная кривой нормального распределения, лежит между $M - z\sigma$ и $M + z\sigma$ (см. Приложение 3, табл. А).

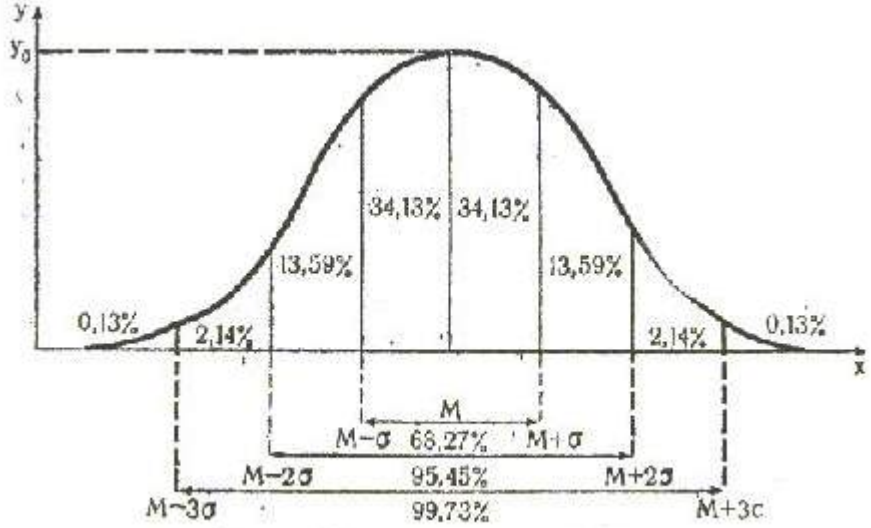


Рис. 16. Нормальное распределение

Так, для $z = 2$ эта площадь равна, как указывалось, 95,45%; для z = 2,5 – 98,76% и т.п. На рис.16 показано, какие доли площади, ограниченной кривой и осью абсцисс, заключены между соседними ординатами (например, между M и $M + \sigma$ 34,13% общей площади).

Распределение примера 4 очень близко к нормальному, в этом легко убедиться непосредственно. Так как σ = 6, то если бы распределение было точно нормальным, 99,72% наблюдений заключены были бы между 147,5 и 183,5. Легко видеть, что здесь ... 99,8%!

Завершая рассмотрение *M, Me* и *Mo,* сделаем существенное замечание. Как мы видели, среднее арифметическое совокупности, состоящей из нескольких групп, может быть выражено как средневзвешенное групповых средних арифметических. Этим свойством, однако, не обладают ни медиана, ни мода: *Me* и *Mo* для групп, из которых состоит изучаемая совокупность, мы ничего не можем сказать о *Me* и *Mo* этой совокупности: ее параметры нужно

[49]

вычислять заново. Таким образом, *Me* и *Mo* не поддаются арифметическим операциям.

Существуют и другие виды средних величин. Поскольку они не получили широкого применения в социологии, ограничимся кратким знакомством с ними.

Средней геометрической величин $x_1, x_2, \ldots, x_N$ по определению, называется величина $G_N = \sqrt[N]{\prod_{i=1}^{N} x_i}$

Если варианты повторяются, то $G_N = \prod_{i=1}^{R} x_i^{Vi}$

Средней гармонической называется величина $H_N = \frac{N}{\sum_{i=1}^{K} \frac{1}{x_i}}$.

Если варианты повторяются, то $H_N = \frac{1}{\sum_{i=1}^{K} \frac{v_i}{x_i}}$. Средняя квадратическая $S_N = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2}$.

Можно доказать, что $H_N \le G_N \le M_N \le S_N$. Оказывается, что H_N, G_N, M_N, S_N могут быть определены с помощью одной формулы: $\alpha_z = \sqrt[z]{\frac{1}{N} \sum_{i=1}^{N} x_i^z}$. Полагая, что z = -1, 0, 1, 2, мы получим H_N, G_N, M_N, S_N соответственно. Доказательство этого составляет содержание *упражнения 14*. Примечание. В случае z=0 нужно сперва вычислить $\ln a_z$, а затем перейти к пределу, когда z→0.

Завершая рассмотрение, отметим, что H_N, G_N, M_N, S_N в отличие от *Mo* и *Me* (две последние величины называют структурными средними) зависят от всех значений признака.

4. Меры вариации

В §3 мы уже познакомились с такими мерами колеблемости, как вариационный размах и дисперсия. Ввиду особой значимости для статистики понятия дисперсии остановимся на нем подробнее.

По определению, дисперсия представляет собой среднее арифметическое квадратов отклонений вариантов от среднего арифметического значения признака для данной совокуп-

[50]

ности, т.е.

$$D = \frac{1}{N}\sum_{i=1}^{N}(x_i - M)^2 = \frac{1}{N}\sum_{i=1}^{k}N(x_i)(x_i - M)^2 \quad \text{(I, 4,1)}$$

Пример 5. Вычислим *M, D,* σ и C_v для квалификации рабочих Одесского судоремонтного завода образованием 7 классов. Для этой совокупности вариационный ряд имеет следующий вид:

Квалификация (в разрядах)	x_1=1	x_2=2	x_3=3	x_4=4	x_5=5	x_6=6	*Всего*
Частота	8	40	42	65	77	53	285

$$M = \bar{x} = \frac{1\cdot 8 + 2\cdot 40 + 3\cdot 42 + 4\cdot 65 + 5\cdot 77 + 6\cdot 53}{285} = 4{,}13 \text{ (разряда)}$$

$$D = \frac{(1-4.13)^2\cdot 8 + (2-4.13)^2\cdot 40 + (3-4.13)^2\cdot 42 + (4-4.13)^2\cdot 65 + (5-4.13)^2\cdot 77 + (6-4.13)^2\cdot 53}{285} =$$

$$= 1.93$$

σ = 1.39 разряда

C_v = 33,8%

Упражнение 15. Найти *M, D,* σ, C_v для рабочих, имеющих общее среднее образование, если распределение имеет вид:

x_i	1	2	3	4	5	6
$N(x_i)$	53	232	212	153	99	34

Ответ: *M* = 3,15; *D* = 1,712,

σ = 1,31 C_v = 41,6%

Так как эти данные получены в одном и том же конкретном социологическом исследовании, целесообразно их сопоставить. Интересно, что у рабочих-судоремонтников с образованием 10-11 классов средний квалификационный разряд на единицу меньше, чем у рабочих с образованием 7 классов. Дело в том, что у рабочих со средним образованием значительно меньше средний стаж (примерно на 9 лет), а для данной специальности стаж в большей мере влияет на квалификацию, чем образование. Подчеркнем, что это локальный вывод, существуют профессии, где решающую роль в квалификации играет образование. Обратим внимание и на то, что группа рабочих с общим средним образованием несколько более разнородна по своему составу в смысле

[51]

квалификации (ср. коэффициенты вариации), а также в смысле стажа и возраста рабочих.

Познакомимся с основными свойствами дисперсии.

1. Если все варианты увеличить (или уменьшить) в одно и то же число, скажем, α раз, то *D* увеличится (или соответственно уменьшится) в α^2 раз.

Мы совершаем переход $x_i \to x_i' = ax_i$. При этом, очевидно, $M = \bar{x} \to \bar{x'} = ax$, а $D \to D' = \alpha^2 D$.

Заметим, что $\sigma \to \sigma' = \alpha\sigma$.

2. Увеличение (или уменьшение) всех вариант на одну и ту же постоянную величину *c* не изменит дисперсию. Теперь, $x_i \to x_i' = x_i + c$, очевидно, $\bar{x} \to \bar{x'} = \bar{x} + c$, а $x_i' - \bar{x'} = x_i - \bar{x}$, т.е. $D'' = D$.

3. При увеличении (или уменьшении) всех частот в одно и то же число раз дисперсия не изменится.

4. Дисперсия относительно средней арифметической равна дисперсии относительно произвольной постоянной за вычетом квадрата разности средней арифметической и этой постоянной.

Представляя $(x_i - \bar{x})$ в виде $\left[(x_i - c) - (\bar{x} - c)\right]$, имеем:

$$D = \frac{1}{N}\left[\sum_i N(x_i)(x_i - c)^2 - 2(\bar{x} - c)\sum_i (x_i - c)N(x_i) + (\bar{x} - c)^2 N\right] = \frac{1}{N}\sum_i N(x_i)(x_i - c)^2 - (\bar{x} - c)^2$$

или: $D(\bar{x}) = D(c) - (\bar{x} - c)^2$.

Отсюда: $D(c) - D(\bar{x}) + (\bar{x} - c)^2$ (I,4,2)

т.е $D(c) \geq D(\bar{x})$

Таким образом, дисперсия относительно среднего арифметического (мы ее будем называть собственно дисперсией, или для простоты, просто дисперсией) обладает свойством минимальности: она меньше дисперсии относительно любой другой величины.

5. Дисперсия равна средней арифметической квадратов вариантов, уменьшенной на квадрат средней арифметической.

[52]

В самом деле, полагая в (I,4,2) $c = 0$, получим:

$D = \overline{x^2} - \bar{x}^2$ (I, 4,3)

Следствием свойств 1 и 4 является равенство

$$D = \frac{\alpha^2}{N}\sum_{i=1}^{k} N(x_i)\left(\frac{x_i - c}{\alpha}\right)^2 - (\bar{x} - c)^2 \qquad \text{(I,4,4)}$$

которое может быть использовано для упрощения вычисления дисперсии.

Упражнение 16. Вернемся к примеру 4 §3 и завершим его рассмотрение, вычислив *D,* σ *и* C_v для распределения по росту взрослых мужчин. Для этого могут быть использованы данные таблицы 9, в седьмой колонке которой приведены величины, необходимые для применения формулы (I,4,4). Как и ранее, c = 165,5, а α = 3. Ответ: D = 36,58; σ= 6,05; C_v = 3,8%. Если читатель выполнит вычисление дисперсии и по формуле (I,4,1), то он сумеет оценить преимущество (I,4,4).

Познакомимся с правилом сложения дисперсий. Будем считать, что изучаемая совокупность разбита на s непересекающихся групп.

Пусть в *r*-ой группе x_i встречается P_{r_i} раз, ясно, что $\sum_{i=1}^{k} P_{r_i} = N_r$, т.е. числу индивидов в *r*-ой группе, а $\sum_{r=1}^{s} P_{r_i} = N(x_i)$ – общему числу индивидов с $x = x_i$. Групповое среднее $\overline{x_r} = \frac{1}{N_r}\sum_{i=1}^{k} x_i P_{r_i}$, а групповая дисперсия суть $(r = \overline{1,s})$:

$$\sigma_r^2 = \frac{1}{N_r}\sum_{i=1}^{k}(x_i - \overline{x_r})^2 P_{r_i} = \frac{1}{N_r}\sum_{i=1}^{k} x_i^2 P_{r_i} - \overline{x_r^2} \qquad \text{(I,4,5)}$$

Межгрупповой дисперсией, по определению, называется средняя арифметическая величина квадратов отклонений групповых средних $(\overline{x_r})$ от общей средней $\bar{x}$, т.е.

$$\partial^2 = \frac{1}{N}\sum_{r=1}^{s}(\overline{x_r} - \bar{x})^2 N_r = \frac{1}{N_r}\sum_{r=1}^{s}\overline{x_r^2} N_r - \overline{x^2} \qquad \text{(I,4,6)}$$

Средняя арифметическая групповых дисперсий:

$$\overline{\sigma^2} = \frac{1}{N}\sum_{r=1}^{s} N_r \sigma_r^2 \qquad \text{(I, 4,7)}$$

[53]

Теперь мы получили возможность сформулировать правило сложения дисперсий:

$$\sigma^2 = \overline{\sigma^2} + \delta^2 \qquad \text{(I, 4, 8)}$$

Покажем, что это действительно так.

Из (I, 4, 5):

$$\sum_{i=1}^{k} x_i^2 P_{r_i} = N_r \sigma_r^2 + \overline{x_r^2} N_r \quad (r = \overline{1, s})$$

Запишем s таких равенств и сложим их почленно, тогда:

$$\sum_{i=1}^{k} x_i^2 N(x_i) = \sum_{r=1}^{s} N_r \sigma_r^2 + \sum_{r=1}^{s} N_r \overline{x_r^2} \qquad \text{(I, 4, 9)}$$

Разделив обе части равенства (I, 4, 9) на N и вычтя из них по $\overline{x^2}$, получим с учетом (I, 4, 5-7): $\sigma^2 = \overline{\sigma^2} + \delta^2$, что и требовалось.

Пример 6.

Пусть совокупность из $N = 150$ индивидов состоит из трех групп (цехов), в первой $N_1 = 40$, во второй $N_2 = 50$, в третьей $N_3 = 60$ человек (здесь s = 3; r = I, 2, 3).

Эмпирические данные сведем в таблицу 10.

Найдем сперва $\overline{x_r} : \overline{x_1} = \dfrac{65 \cdot 7 + 75 \cdot 12 + 85 \cdot 15 + 55 \cdot 6}{40} = 80$ (руб.)

Аналогично: $\overline{x}_2 = 95$ руб.; $\overline{x_3} = 105$ руб.

В качестве *упражнения 17* предлагаем вычислить $\overline{x}$ сначала как средневзвешенное $\overline{x_2}$, а затем непосредственно, по определению. В обоих случаях должен получиться один и тот же результат: $\overline{x} = 95$ руб.

Далее. Найдем σ_r (*r*= I, 3, 2). Например,

$$\sigma_1^2 = \frac{(65-80)^2 \cdot 7 + (75-80)^2 \cdot 12 + (85-80)^2 \cdot 15 + (95-80)^2 \cdot 6}{40} = 90$$

Упражнение 18. Найти σ_2^2 и σ_3^2. Ответ: 140; $66\frac{2}{3}$.

Следующий шаг. Вычислим межгрупповую дисперсию:

$$\delta^2 = \frac{1}{N} \sum_{r=1}^{s} (\overline{x_r} - \overline{x})^2 N_r = 100,$$

а также $\overline{\sigma^2} = 97\frac{1}{3}$ и $\sigma^2 = 197\frac{1}{3}$

Таким образом, $\sigma^2 = \delta^2 + \overline{\sigma^2}$.

[54]

Приведем примеры вычислений M, σ и C_v для двухмерных распределений.

Пример 7. Пусть первый признак X – заработная плата рабочих (в рублях), а второй – Y – квалификация (в разрядах). Второй признак дискретный, а первый интервальный (величины соответствующих интервалов представлены в табл. II). 2131 рабочий, подвергнутые обследованию, в частности, по признакам X и Y распределились так, как

Таблица 10

Пример расчета межгрупповой и внутригрупповой дисперсии

Заработная плата, руб.	x_i, руб.	P_{1i}	P_{2i}	P_{3i}	$N(x_i)$
60—70	65	7	1	0	8
70—80	75	12	5	0	17
80—90	85	15	9	4	28
90—100	95	6	18	8	32
100—110	105	0	12	32	44
110-120	115	0	5	16	21
Всего	—	40	50	60	150

показано в табл. 11. Например, 18 человек имеют первый разряд и получают до 80 руб., 28 – первый разряд и зарплату в интервале от 80 до 100 руб. Всего рабочих с первым разрядом 121, со вторым 523 и т.д. Всего получающих зарплату до 80 руб. – 107 чел., от 80 до 100 руб. – 216 и т.д.

Таким образом, в колонке $N(x_i)$ по сути представлено распределение рабочих по разрядам, а в строке $N(y_i)$ – по заработной плате. Это вариационные ряды типа ранее рассмотренных. Кроме того, наша табл. 11 содержит специфические ряды типа: распределение рабочих с данной зарплатой по разрядам и распределение рабочих с данным разрядом по величине заработной платы.

В столбцах приведены также средние значения X (например, средняя зарплата рабочих с первым квалификационным разрядом 113,1 руб., а средний разряд рабочих, заработная плата которых от 180 до 200 руб., составляет 4,03 разряда). Далее представлены соответствующие σ и C_v.

Мы приводим эту таблицу не столько из-за ее информационной ценности, сколько для того, чтобы читатель мог

[55]

Таблица 11

Пример расчета M, σ, C_v **(зарплата)**

разряд	Зарплата, руб.												
	до 80	80-100	100-120	120-140	140-160	160-180	180-200	200-220	свыше 220	$N(x_i)$	$\bar{x}$	σ_x	$C_v^{(x)}$
1	18	28	26	26	15	6	1	1	0	121	113,1	30,44	26,9
2	42	77	128	140	67	31	16	13	9	523	125,6	34,93	27,8
3	33	50	84	139	123	20	19	11	8	527	134,4	33,80	25,1
4	7	45	71	66	65	72	44	24	22	416	148,6	42,35	28,4
5	4	12	49	50	57	46	34	38	54	344	167,3	48,51	29,0
6	3	4	23	53	47	29	12	9	20	200	155,7	41,58	26,6
$N(y_i)$	107	216	381	474	374	244	126	96	113	2131			
$\bar{y}$	2,49	2,75	3,15	3,28	3,59	3,85	4,03	4,16	1,60	–			
σ_y	1,13	1,18	1,39	1,41	1,37	1,27	1,17	1,22	1,10	–			
$C_v^{(y)}$	45,3	43,0	44,3	43,0	38,2	33,0	29,0	29,3	23,9	–			

при желании проверить себя и рассчитать показатели, которые были рассмотрены в предыдущих параграфах.

Что же касается содержательной интерпретации данных, кроме достаточно очевидных утверждений типа «с увеличением разряда увеличивается средняя заработная плата», из нее можно почерпнуть менее очевидное: с увеличением заработной платы группы рабочих становятся все более однородными по уровню квалификации (монотонное уменьшение C_v), хотя с увеличением разряда вариация заработной платы не изменяется: разброс примерно один и тот же.

Пример 8.

При изучении связи между признаками квалификация *(X)* и удовлетворенность специальностью (*Y*), в частности, была получена такая таблица (таблица 12).

Как и ранее, *X* выражается в разрядах $(x_i = i, i = \overline{1,6})$. Признак *Y* – качественный, его позиции: «удовлетворен», «не знаю, трудно сказать», «не удовлетворен» обозначены в таблице соответственно y_1, y_2, y_3. Удовлетворенность группы рабочих описывается с помощью индекса $J_{спец} = \dfrac{N_+ - N_-}{N_+ + N_0 + N_-}$, где N_+ – число удовлетворенных, N_- – неудовлетворенных, N_0 – не выразивших определенное отношение.

Так как в дальнейшем нам придется неоднократно рассматривать индексы для группы, отметим, что *J* принимает значения, заключенные между –1 и 1, причем –1 соот-

ветствует случаю, когда все работники не удовлетворены, 1 означает, что все удовлетворены, а 0 получается в случае, когда число удовлетворенных специальностью равно числу неудовлетворенных. Аналогично конструируются индексы удовлетворенности работой, различными элементами рабочей ситуации и т.д.

Возвратимся к табл. 12. Кроме «очевидных» утверждений типа «с увеличением квалификации увеличивается и удовлетворенность специальностью», из нее следует, что группы индивидов с разной удовлетворенностью специальностью примерно одинаковы по вариации квалификации, а с увеличением квалификации резко возрастает однородность групп по степени удовлетворенности: последняя складывается из все более согласованных индивидуальных оценок.

До сих пор мы рассматривали упорядоченные (количественные и качественные) признаки. Возникает вопрос, что может служить мерой вариации классификационных признаков?

Вариация классификационных признаков. Очевидно, меры, разработанные для признаков, значения которых числа, оказываются теперь непригодными: между объектами разных классов нет упорядочения (все классы равноправны – нельзя выделить континуум, в котором можно было бы упорядочить национальную или расовую принадлежность, членство в различного рода организациях или причины

[57]

увольнения с предприятия и т.д.), нет нуля, нет интервалов, теряют смысл такие понятия, как диапазон, размах, отклонение, столь привычные и удобные, когда значения признаков – числа.

Тем не менее объекты, входящие в разные классы, обладают различными качествами в смысле изучаемого признака: у них разный пол и разная национальность, они принадлежат к разным организациям или указывают разные причины увольнения и т.д.

Таблица 12

Пример расчета M, σ **и** C_v **(удовлетворенность)**

X	Y			$N(x_i)$	$\bar{x}$	σ_x	C_v^x (%)
	y_1	y_2	y_3				
1	66	4	30	100	0,36	1,09	302,8
2	327	16	77	420	0,59	0,86	145,8
3	353	25	61	439	0,66	0,76	115,2
4	295	25	34	354	0,74	0,65	87,8
5	271	16	15	302	0,85	0,49	57,6
6	172	3	6	181	0,92	0,38	41,3
$N(y_i)$	1484	89	223	1796			
$\bar{y}$	3,60	3,47	2,75				
σ_y	1,51	1,22	1,24				
C_v^y (%)	41,9	35,2	45,1				

Попробуем оценивать вариацию с помощью различия в качестве. Чем больше число различных пар объектов, тем, очевидно, больше вариация. Допустим, что у нас всего 2 класса объектов A и B (например, признак «пол»), численность которых N_A и N_B соответственно (объем совокупности $N = N_A + N_B$). В этом случае число различных пар объектов $N_A \cdot N_B$ (скажем, каждый мужчина, очевидно, отличается от каждой женщины: на одного приходятся N_B женщин, т.е. N_B различий, а на всех N_A мужчин $N_A \cdot N_B$ различий). Для того, чтобы сконструировать нормированную меру, определим, в каком случае число пар максимально.

Как известно, среднее геометрическое двух чисел a и b не превосходит среднего арифметического и равно ему, если $a = b : \sqrt{ab} \le \frac{a+b}{2}$.

Пусть $a = N_A^2, b = N_B^2$, тогда имеем $N_A N_B \le \frac{N_A^2 + N_B^2}{2}$,

[58]

следовательно, $(N_A N_B)$ максимально, когда численности классов одинаковы, т.е. равны $\frac{N}{2}$; $(N_A N_B)_{\max} = \frac{N^2}{4}$, а искомая мера $\frac{4N_A N_B}{N^2} = \left(\frac{G_2}{M_2}\right)^2$. Итак, вариация максимальна, когда классы равнонаполненные, она при этом равна 1. Вариации нет, если, скажем, $N_A = 0$ ($N = N_B$ – все объекты однотипны), мера вариации при этом, очевидно, обращается в нуль.

А как быть, если классов больше чем 2, например, 3? Для двух классов мера равна квадрату отношения среднего геометрического к среднему арифметическому численностей классов. Казалось бы, в случае трех классов A, B, C, мера вариации должна быть $(G_3 / M_3)^2 = \frac{9N_A N_B N_C}{N^3}$. Легко видеть, что это не так. Допустим, что $N_A = 0$, тогда величина $(G_3 / M_3)^2$ обращается в нуль, хотя совокупность неоднородна: остались объекты типа B и C. Как же быть?

Составим величину

$$\alpha_3 = \frac{N_A N_B + N_A N_G + N_B N_G}{3\left(\frac{N}{3}\right)^2} \qquad \text{(I, 4,10)}$$

Она обращается в нуль, если по крайней мере два класса пусты (скажем, $N_A = N_B = 0$, т.е. совокупность однородна, состоит только из объектов типа C).

Максимальное значение обсуждаемая величина принимает при $N_A = N_B = N_C$, которое, как легко видеть, равно 1 (при этом различия максимальны). Величину α_3 можно принять в качестве меры вариации.

Упражнение 19. Показать, что $ab + ac + bc \le \frac{(a+b+c)^2}{3}$, причем равенство достигается при a=b=c. Указание: использовать трижды – для всех пар – неравенство между G_2 и M_2. Итак, $0 \le a_3 \le 1$, причем нуль соответствует однородной совокупности (отсутствие вариации), а единица – максимально неоднородной (максимальная вариация, случай равнонаполненных классов).

Упражнение 20. Рассмотреть случай $k = 4 (N = N_1 + N_2 + N_3 + N_4)$

Ответ:

$$\alpha_4 = \frac{8}{3}\frac{N_1 N_2 + N_1 N_3 + N_1 N_4 + N_2 N_3 + N_2 N_4 + N_3 N_4}{N^2} \qquad \text{(I, 4,11)}$$

[59]

Рассмотрим общий случай (произвольное k). Теперь число различий $A = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_i N_j$.

Найдем максимальное А, которое соответствует случаю $N_l = \frac{N}{k} (l = \overline{1,k})$:

$$A_{\max} = \frac{N_2}{k^2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 1 = \frac{N^2}{k^2} \left(\sum_{j=2}^{k} 1 + \sum_{j=3}^{k} 1 + \ldots + \sum_{j=k-1}^{k} 1 + \sum_{j=k}^{k} 1 \right) = \frac{N^2}{k^2} \left[(k-1) + (k-2) + \ldots + 2 + 1 \right] =$$

$$= \frac{N^2 (k-1)}{2k},$$

таким образом,

$$\alpha_k = \frac{2k}{k-1} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} v_i v_j \qquad \text{(I, 4,12)}$$

Для описания вариации можно использовать также и энтропийную меру (см. § 5 главы II).

Квантили. Медиана, как мы видели, это значение признака, которое обладает таким свойством: 50% вариантов меньше, чем Me, 50% – больше. Естественным обобщением медианы является понятие квантиля. Квантиль делит сумму частот на заданное число равных частей. Число частей может быть различным, отсюда и разные квантили – квартили, децили, перцентили.

Квартиль. Квартиль (Q_i) делит сумму частот на четыре равные части. Очевидно, квартилей всего три: Q_1, Q_2, Q_3; Q_1 например, это значение признака, которое обладает таким свойством: 25% вариантов меньше, а 75% – больше его. Q_2 это Me, а Q_3 – значение признака, 75% вариантов меньше которого, а 25% – больше.

Прямые $x = Q_i (i = 1,2,3)$ делят площадь, ограниченную кривой распределения на 4 равные части: $S_1 = S_2 = S_3 = S_4$

На рис. 17а изображено распределение, а на рис. 17б показаны квартили на графике кумулятивной кривой. Подчеркнем, что точки, соответствующие квартилям, вообще говоря, делят отрезок $[x_{\min}, x_{\max}]$ на четыре неравные части. Между Q_1 и Q_3 заключена половина всех вариантов. Чем более плотно распределение, тем отрезок $[Q_1, Q_3]$ меньше. Таким образом, своеобразной мерой «разброса» может служить величина $\Delta Q = Q_3 - Q_1$.

[60]

Дециль. Дециль (D) делит сумму частот на 10 равных частей. Всего децилей, очевидно, девять: $D_1, D_2, ... D_9$. Ясно, что $D_5 = Q_2 = Me$. В качестве меры разброса используется также величина $\Delta D = D_9 - D_1$.

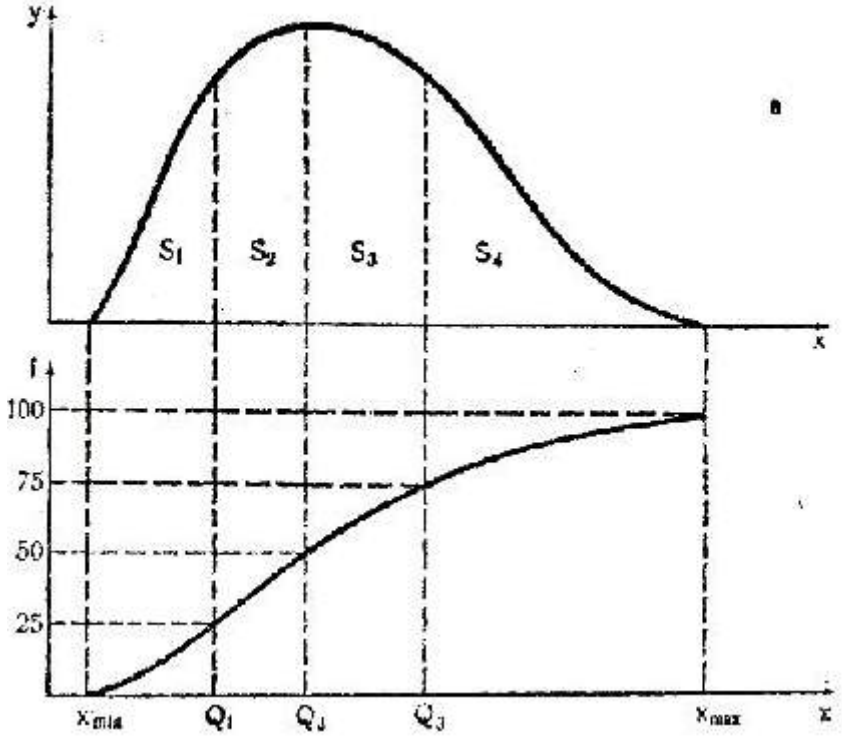


Рис. 17. Квартили на графике распределения (а) и на графике кумулятивной кривой (б)

Перцентиль, по определению, делит сумму частот на 100 равных частей: $C_1, C_2, \ldots, C_{99}$ Легко видеть, что, например, $D_1 = C_{10}$, $Q_1 = C_{25}$, $Me = C_{30}$, $Q_3 = C_{75}$ и т.д.

Как вычислять квантили в случае интервальных рядов? Вспоминая вывод формулы для $Me(Q_2)$, легко понять, что

$$Q_1 = x_i + I_i \frac{0{,}25N - F_{l-1}}{N_l},$$

$$Q_3 = x_i + I_l \frac{0{,}75N - F_{l-1}}{N_l},$$

где l – номер интервала, в который попадает соответствующий квантиль.

[61]

Упражнение 21. Вывести формулы для Q_1 и Q_3. Аналогично, например,

$$D_3 = x_l + I_l \frac{0{,}3N - F_{l-1}}{N_l}$$

$$C_{99} = x_l + I_l \frac{0{,}99N - F_{l-1}}{N_l} \text{ и т.д.}$$

Отметим, что квантиль – мера, применимая к самым различным типам упорядоченных данных. При вычислении квантилей вместо частот можно использовать частости.

Пример 9. По данным таблицы № 8 рассчитать Q_1 для годового семейного дохода в США (1959 г.). Нетрудно видеть, что $l = 2$, $x_l = x_2 = 2000$, $f_{l\text{-}1} = f_1$ 14, $v_l = v_2 = 21$, теперь $Q_1 \approx 3050$. Таким образом, 25% семей имели доход, меньший 3050 дол.

Упражнение 22. Вычислить Q_3, ΔQ, D_9. Нередко частоты крайних вариантов очень малы, величина вариационного размаха может создать впечатление большей колеблемости (величины вариации), нежели та, которая наиболее характерна для изучаемого распределения. В таких случаях целесообразно вычислять ΔQ или ΔD, в которых отражен диапазон, включающий в себя соответственно 50% и 80% всех наблюдений.

Упражнение 23. Какой процент американских семей имел доход ниже прожиточного минимума (3000 дол.)?

Далее мы рассмотрим применение изученных величин (Me, Q_i, ΔQ) к одной социологической задаче – измерению установки индивидов.

Пример 10. Шкала Терстоуна. С помощью этой шкалы измеряется ориентация (отношение, установка). Терстоун непосредственно изучал отношение к церкви (дальше мы подробно рассмотрим соответствующую процедуру), однако предложенный способ может быть использован для измерения различных установок. Итак, изучаемый признак – отношение.

Пункты шкалы устанавливаются не произвольно, а с помощью отбора суждений, осуществляемого судьями. Сперва при участии представителен обследуемого массива был составлен список, содержавший более ста высказываний, отражающих различное отношение к изучаемому феномену. Затем 300 судьям, представлявшим модель исследуемой аудитории, было предложено разложить карточки с высказываниями на 11 кучек: в первой должны быть суж-

дения наиболее благоприятные для церкви, во-второй – менее и т.д. до 11-ой, куда попадают наименее благоприятные суждения.

После того, как судьи завершили работу, нужно установить цену каждого суждения, меру согласованности судейских решений по каждому суждению и отобрать набор суждений, с помощью которых исследователь может изучать рассматриваемое отношение индивидов данной общности.

Цена суждения определялась как медиана распределения судейских решений, мера согласованности – квантильное отклонение.

Чтобы обработать результаты работы судей, для каждого из суждений первоначального списка составляется такая таблица:

Пункты шкалы	N_i(число судей, поместивших данное суждение в этот пункт)	v_i (% к общему числу судей)	f_i (кумулятивный %)
1	2	3	4
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	12	4%	4%
7	12	4%	8%
8	60	20%	28%
9	66	22%	50%
10	90	30%	80%
11	60	20%	100%
	300	100%	

Затем строится кумулята (рис. 18). При этом предполагается, что отношение изменяется континуально, пункты 1, 2, ..., 11 – отдельные точки, которые выделяют в данном континууме интервалы; ординаты кумуляты соответствуют серединам соответствующих интервалов.

Для представленного на графике суждения, как видно из чертежа, Me = 8,5; Q_1=7,3, Q_3=9,3, ΔQ=2,0. Проделав такую процедуру со всеми суждениями, в итоговую шкалу отбирают те, которые: 1) покрывают более или менее равномерно всю шкалу; 2) имеют наиболее согласованные

[63]

оценки, т.е. из нескольких суждений с близкими *Me* предпочтение отдается суждению с минимальным квартильным отклонением ΔQ.

Окончательная шкала содержит 10–15 суждений, каждое из которых имеет свой «вес» (цену) – медиану судейских решений. Отобранные суждения предлагаются респон-

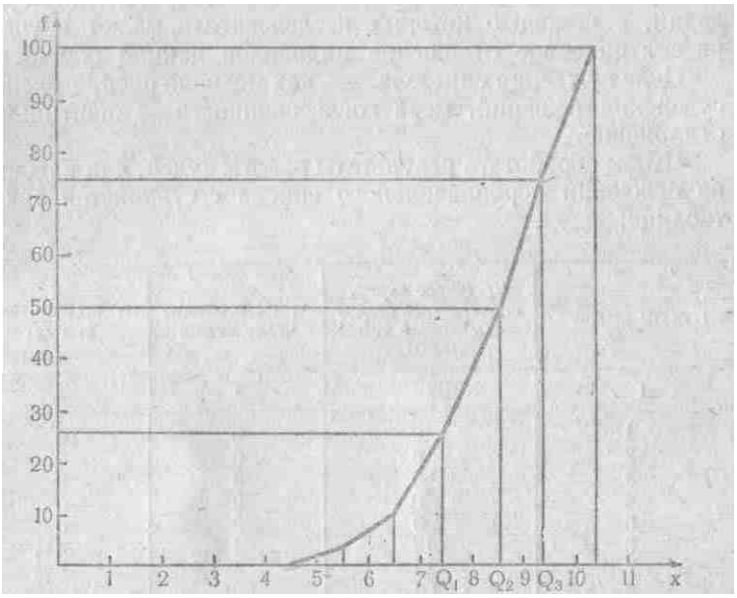


Рис. 18. Кумулята для построения шкалы Терстоуна

денту. Его ранг по данной шкале – медиана «весов» принятых им суждений, т.е. суждений, с которыми он согласен. Если респондент А согласен с такими пятью суждениями, у которых «веса»: 4,4; 4,8; 5,1; 5,6; 6,1, то его ранг 5,1. Если респондент В выбрал четыре суждения с «весами»: 7,6; 8,1; 8,5; 8,7, то его ранг 8,3 (медиана «весов» в случае четного числа суждений, по определению, $\frac{8{,}1 + 8{,}5}{2} = 8{,}3$).

Отметим, что шкала Терстоуна обладает рядом недостатков, устраненных в более совершенных методах[35].

[64]

[35] *Клигер С. А., Косолапов М. С., Толстова Ю. Н.* Шкалирование при сборе и анализе социологической информации. М., 1978, с. 71—81.

Глава II

КОРРЕЛЯЦИИ

1. Функциональная и корреляционная зависимости. Корреляционные таблицы. Критерий Пирсона

Если данному значению одной величины соответствует вполне определенное значение другой, то говорят, что между этими величинами имеет место функциональная зависимость. Такого рода зависимость, например, имеет место между силой гравитационного взаимодействия двух масс m_1 и m_2 и расстоянием r между ними; $F = \gamma \frac{m_1 \cdot m_2}{r^2}$, где γ — гравитационная постоянная (закон Ньютона).

Функционально связаны: общий стаж работы Y и стаж работы на данном предприятии X (здесь $Y=aX+b$, где b — стаж работы до поступления на это предприятие, a обычно равно 1; если же год работы засчитывается, скажем, за 2, то a=2 и т.д.); выработка и время работы определенного рабочего (в последнем примере связь может носить довольно сложный характер и ее трудно будет описать аналитически, в таком случае ее можно отобразить графически).

Однако далеко не всегда зависимость может иметь столь простой (или относительно простой) характер. Часто случается так, что определенному значению одной величины соответствует целый комплекс значений другой, представляющий собой ряд распределения, причем при изменении данной величины меняется ряд распределения и его среднее. В таких случаях говорят о *корреляционной зависимости.* Она отражает тенденцию возрастания (положительная корреляция) или убывания (отрицательная корреляция) одной переменной величины при возрастании другой.

Классический пример такого рода зависимости — связь между ростом отцов (*X)* и детей (*У*). Конечно, у высокого отца может быть низкорослый сын, а у низкорослого — высокий, но в совокупности случаев прослеживается тенденция увеличения *Y* с увеличением *X*, т.е. Положительная

[65]

корреляция. Если каждую пару значений этих величин изобразить на плоскости в прямоугольной системе координат с помощью точек, то наносимые точки не расположатся на одной кривой, как в случае функциональной связи (рис. 19а, где каждому x_i, например, соответствует вполне определенное y_i на кривой), а образуют некоторое «облако», называемое корреляционным полем (рис. 196). В нашем примере это облако не окажется абсолютно бесформенным, оно вы-

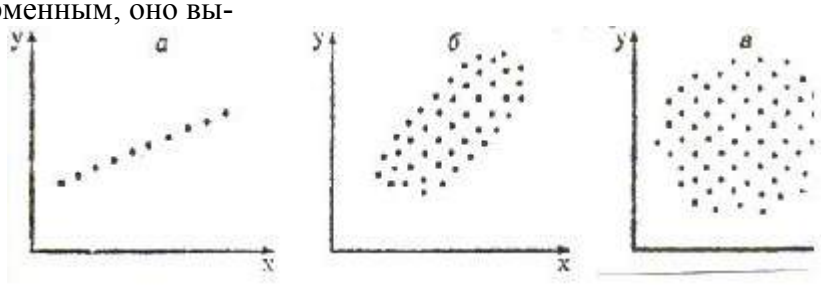


Рис. 19. Корреляционное поле для различных видов связи: а — функциональная связь; б—корреляционная связь; в— отсутствие связи.

тянется так, что будет прослеживаться увеличение среднего *У* с увеличением *X*.

Корреляционная зависимость имеет место также между количеством удобрений и урожайностью, размером предприятий и себестоимостью, спросом на товары и ценой на рынке и т.д.

Корреляционная зависимость не является абсолютно точной, полной. В ней отражается множественность причин и следствий. Каждое явление находится под влиянием большого числа причин, действующих с разной силой. Изучая влияние *X* на *Y,* мы выделяем один фактор, но на данный признак *Y* оказывают влияние и многие другие, что обусловливает корреляционный характер зависимости.

Например, станем рассматривать влияние стажа на производительность труда рабочего. Ясно, что стаж влияет на производительность, но не может определять ее полностью» так как на производительность влияют квалификация и образование, возраст и состояние здоровья и другие факторы. Таким образом, стаж далеко не единственный фактор производительности, связь между этими переменными корреляционная. И вообще: в силу сложности, многофакторности общественной жизни связи между социальными переменными практически всегда корреляционные.

Функциональная и корреляционная связи могут быть, а могут не быть причинно-следственными. Логическая природа рассматриваемых «сечений» (функциональная — кор-

[66]

реляционная и причинно-следственная — не причинно-следственная) принципиально различна.

Рассмотрим пример. Как известно, между давлением P, объемом V, абсолютной температурой T и массой газа M существует функциональная зависимость

$PV = CMT$

(здесь C — константа)

Четыре величины P, V, M, T связаны функционально и вопрос о том, какая из них причина, какая следствие в общем случае лишен смысла. Однако в конкретной физической ситуации он может быть правомерным. Допустим, что данная масса газа находится под постоянным давлением. (Сосуд закрыт поршнем с определенным «гнетом»). Начинаем нагревать сосуд. С увеличением T будет увеличиваться V, причем каждому T_i соответствует свое вполне определенное V_i. Значит, в случае функциональной зависимости такого рода причиной является нагревание, следствием — расширение объема. В упрощенной ситуации (при абстрагировании от ряда явлений, что часто законно) можно говорить о причинной зависимости между одной причиной и одним следствием.

В случае корреляционной связи все значительно сложнее. Здесь, как уже подчеркивалось, имеет место множественность причин: любое явление находится под влиянием большого числа факторов, каждый из которых имеет, вообще говоря, различную «силу». Наличие корреляции свидетельствует, что либо одно из двух выделяемых явлений есть частичная причина другого, либо оба явления — следствие общих причин. При этом «статистик, как таковой, будучи вполне компетентным в установлении корреляции между любыми величинами, к какой бы области они ни принадлежали, не компетентен в высказывании причинных суждений. Для этого мало быть статистиком, а нужно быть биологом, медиком, метеорологом, экономистом и т.д., смотря по области исследования»[1]. Таким образом, установление корреляции еще не служит само по себе показателем существования причинно-следственной связи.

Чтобы проиллюстрировать эту мысль, приведем, на наш взгляд, показательный пример[2].

[67]

[1] *Слуцкий Е. Е.* Теория корреляции и элементы учения о кривых распределения. Киев, 1912, с. 133.

[2] Заимствован из книги: *Richardson C.H.* An introduction to statistical analysis. New York, 1949, p. 268—269.

Пример 11. Для признаков *X* и *Y,* задаваемых таблицей 13, коэффициент корреляции (см. § 4 этой главы) $r = 0{,}98$, т.е. между *X* и *Y* есть значимая прямая связь. Здесь: *X* — общая заработная плата школьных работников в миллионах долларов, а *Y* — общее потребление вина и ликеров в США в миллионах галлонов. Едва ли можно утверждать, что заработная плата школьных работников непосредственно зависит от потребления вина и ликеров или потребление винно-ликерных изделий от зарплаты школьных работни-

Таблица 13

Зарплата (*X*) и потребление вина (*Y*) в США с 1870 по 1910 годы

Признаки	Годы								
	1870	1875	1880	1885	1890	1895	1900	1905	1910
X	38	55	56	73	92	114	138	177	254
Y	30	38	51	69	97	114	135	169	205

ков. Высокий коэффициент корреляции означает тесную линейную статистическую связь между двумя переменными и указывает лишь на возможную причинную связь.

Измерение корреляции — это часть проблемы, интерпретация результатов — другая, зачастую более трудная. Обсуждаемую корреляцию можно объяснить, обратившись к истории США. Период с 1870 г. по 1910 г. характеризовался бурным развитием экономики этой страны. Быстро увеличивалось население, развивались торговля, промышленность, сельское хозяйство. Росло число занятых во всех сферах хозяйства, росла и заработная плата (в частности — учителей). Росло потребление вообще (в частности — вин и ликеров).

В исследованиях, осуществленных В. Шубкиным в Новосибирске[3], была установлена корреляционная связь между зарплатой родителей и успеваемостью учеников. Эта связь не является причинно-следственной. Оказывается, существует положительная связь между образованием и зарплатой, очевидна связь между образованием родителей и успеваемостью учеников. Следовательно, и в этом случае связь двух признаков является следствием третьей общей

[68]

[3] Количественные методы в социологии. М., 1966, с. 96.

причины. Связи такого рода иногда называют связями сопутствия.

Таким образом, количественный анализ не может заменить специальные знания, но может сделать теоретическое мышление исследователя более эффективным, так как дает возможность отбросить несущественные связи, очертить круг поисков. Количественный анализ позволяет также

Таблица 14

Зависимость между стажем (*X*) и производительностью труда (*Y*) рабочих промышленного предприятия

X	*Y*						
	y_1=20	y_2=24	y_3=28	y_4=32	y_5=36	y_6=40	$N(x_i)$
X_1=2	9	4	1	0	0	0	14
X_2=6	1	10	9	3	0	0	23
X_3=10	0	2	6	14	6	0	28
X_4=14	0	0	1	10	18	6	35
$N(y_j)$	10	16	17	27	24	6	100

сравнивать влияние различных факторов (частная корреляция).

Перейдем непосредственно к процедурам описания корреляционных связей. Сначала рассмотрим корреляционную таблицу на конкретном числовом примере связи между стажем *X* и производительностью *Y*.

Пример 12. Уже отмечалось, что эта связь не является функциональной: зная стаж рабочего, мы не можем точно указать его производительность. В среднем же, если ограничиться не очень большими *X* (большим *X* соответствует большой возраст и, следовательно, некоторое уменьшение производительности), то увеличению *X* должно соответствовать увеличение *Y* (точнее — среднего значения *Y*). Попытаемся установить вид этой зависимости на примере. Пусть имеются данные о стаже *(X)* и производительности *(Y),* *N*=100 рабочих промышленного предприятия.

Выделим стажные группы с интервалом, например, в 4 года и представим их в корреляционной таблице серединами соответствующих интервалов: x_i = 2, 6, 10, 14 (у нас 4 интервала, в изучаемой совокупности рабочие со стажем

[69]

до 16 лет включительно). Допустим, что производительность измеряется количеством изготовленных деталей, и рабочие могут изготавливать от 18 до 42 деталей за смену. Сгруппируем количество деталей в 6 интервалов. Каждый из них представлен своей серединой y_j = 20, 24, 28, 32, 36, 40. Сведем данные в итоговую корреляционную таблицу (табл. 14).

Как читать ее? Например, в 4 столбце (y_4) 3 строки (x_3) стоит цифра 14. Это значит, что 14 рабочих имеют стаж от

Таблица 15

Общий вид корреляционной таблицы двух признаков.

X	Y						$N(x_i)$
	y_1	y_2	…	y_j	…	y_l	
x_1	N_{11}	N_{12}	…	N_{1j}	…	N_{1l}	$N(x_1)$
x_2	N_{21}	N_{22}	…	N_{2j}	…	N_{2l}	$N(x_2)$
…	…	…	…	…	…	…	…
x_i	N_{i1}	N_{i2}	…	N_{ij}	…	N_{il}	$N(x_i)$
…	…	…	…	…	…	…	…
x_k	N_{k1}	N_{k2}	…	N_{kj}	…	N_{kl}	$N(x_k)$
$N(y_j)$	$N(y_1)$	$N(y_2)$	…	$N(y_j)$	…	$N(y_l)$	N

8 до 12 лет (x_3=10) и производят от 30 до 34 (y_4=32) деталей за смену. Это число естественнее обозначить N_{34}. В последнем столбце ($N(x_i)$) второй строчки стоит цифра 23. Она означает, что всего рабочих со стажем от 4 до 8 лет (x_2=6) 23 чел. Это число мы будем обозначать $N(x_2)$.

В первом столбце (y_1) последней строки стоит цифра 10. Она показывает, сколько всего рабочих изготовляют за смену от 18 до 22 деталей. В наших обозначениях это $N(y_1)$.

Итак, N_{ij}— обозначения внутриклеточных частот, $N(x_i)$ — маргиналов (итогов) по X, $N(y_j)$ — по Y. Саму корреляционную таблицу мы будем для краткости обозначать $\{N_{ij}\}$. В нашем случае $i=\overline{1,4}; j=\overline{1,6}.$ Заметим, что в самом общем случае, когда $i=\overline{1,k}$, а $j=\overline{1,l}$, корреляционная таблица[4] принимает такой вид (табл. 15). Ясно,

[70]

[4] Корреляционная таблица, таблица сопряженности двух признаков, таблица двухмерного распределения («двухмерка»), комбинационная таблица — синонимы (первые два названия чаще используют статистики, остальные — чаще социологи).

сумма всех частот равна: 1) сумме X-маргиналов, 2) сумме *Y*-маргиналов; 3) числу опрошенных:

$$N = \sum_{i=1}^{k} N(x_i) = \sum_{j=1}^{l} N(y_j) = \sum_{i=1}^{k} \sum_{j=1}^{l} N_{ij}$$

Вернемся, однако, к корреляционной таблице для признаков стаж — производительность.

Мы видим, что каждому x_i, соответствует не определенное значение *y*, а *распределение:* y_j, N_{ij} $(j = \overline{1,l})$.

Для x_1:	y_{1j}	20	24	28	
	N_{1j}	9	4	1	
для x_2:	y_{2j}	20	24	28	32
	N_{2j}	1	10	9	3

и т.д.

При изменении *X* меняется распределение *Y*: и сами варианты (при переходе к x_2 появляется вариант 32), и их частоты.

Если внимательно изучить корреляционную таблицу, можно заметить, что с увеличением *X* увеличивается *Y*. Чтобы сделать эту зависимость наглядной, проследим за изменением групповых средних. Для группы $x_1 : \overline{y_1} = \dfrac{(20 \cdot 9 + 24 \cdot 4 + 28 \cdot 1)}{14} = 21{,}7$.

Аналогично для $x_2 : \overline{y_2} = 26{,}4;\ x_3 : \overline{y_3} = 31{,}4;\ x_4 : \overline{y_4} = 35{,}2$.

Упражнение 24. Построить график по точкам $(x_i, \overline{y_i})$.

Из графика видно, что точки лежат почти на одной прямой, т.е. зависимость практически линейная: $\overline{y_1} = ax_i + b$.

Теперь можно дать такое определение корреляционной зависимости: если каждому значению одной величины X*(xi)* соответствует не одно значение, а групповая средняя другой величины $Y(\overline{y_i})$, то зависимость между *X* и *Y* является корреляционной (некоторым значениям *X* при этом, разумеется, может соответствовать лишь одно значение *Y*).

Уравнения, описывающие эту зависимость, называются корреляционными, или регрессионными, а соответствующие им графики — кривыми регрессии.

В рассмотренном примере кривая регрессии — прямая линия. В общем случае зависимость, конечно, не является прямолинейной.

Замечание. Если $\overline{y_1} = \overline{y_2} = ... = \overline{y_k}$, то корреляционной зависимости нет: изменению *X* не сопутствует изменение групповых средних *Y*.

[71]

Распределение объектов по клеткам таблицы, очевидно, зависит от характера связи между признаками. Зададимся вопросом: какой вид должна иметь корреляционная таблица, если связи нет?

Рассмотрим клетку (i, j). Она находится в i-ой строке, на долю которой приходится $N(x_i)$ объектов. Если связи нет, то число объектов в данной клетке будет определяться только общим числом объектов в столбце: чем больше $N(y_j)$, тем больше их окажется и в клетке (i, j), т.е. на ее долю придется $\frac{1}{N} \bullet N(y_j)$ частей $N(x_i)$. Итак, если связи нет, то в (i, j) попадет $\frac{1}{N} N(x_i) \bullet N(y_j)$ объектов. Станем обозначать эту частоту N_{ij}^0 и называть теоретической в отличие от фактически наблюдаемой — эмпирической $N_{ij} : N_{ij}^0 = \frac{1}{N} N(x_i) N(y_j)$.

Какова мера отклонения эмпирической таблицы от теоретической?

Для данной клетки это, конечно, $\Delta_{ij} = N_{ij} - N_{ij}^0$. А для таблицы? Если суммировать Δ_{ij}, то отклонения разных знаков будут компенсироваться и мера различия таблиц получится заниженной. Чтобы избежать этого, нужно «освободить» Δ_{ij} от знаков. Целесообразно перейти к $\Delta^2_{ij.}$

Рассмотрим две клетки: (i, j) и (i', j'), пусть $N_{ij}^0 > N_{i'j'}^0$, а $\Delta_{ij}^2 = \Delta_{i'j'}^2$. В каком случае мера отклонения больше? Очевидно, во втором, так как то же Δ^2 приходится на меньшую частоту. Следовательно, за меру отклонения эмпирической таблицы от теоретической естественно принять, следуя Пирсону, величину

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(N_{ij} - N_{ij}^0\right)^2}{N_{ij}^0} \quad \text{(II,1,1)}$$

Эта мера называется критерием χ^2 («хи-квадрат»), или критерием Пирсона. Заметим, что само обозначение χ^2 подчеркивает неотрицательность критерия; $\chi^2=0$, если все $N_{ij}=N^0_{ij}$; во всех остальных случаях $\chi^2>0$.

В силу разного рода случайных обстоятельств N_{ij} могут отличаться от N^0_{ij} даже в том случае, когда эмпирическое распределение в принципе соответствует теоретическому. Конечно, при этом χ^2 должно быть невелико: большие значения критерия означают принципиальное несоответствие

[72]

обсуждаемых распределений. Каковы же значения χ^2, при которых можно считать, что отклонение $\{N_{ij}\}$ от $\{N^0_{ij}\}$ носит случайный характер?

Так как речь идет о случайных событиях, заключения могут носить лишь вероятностный характер: утверждения о расхождении таблиц высказываются с определенной вероятностью[5], например, с вероятностью p=0,99 или, скажем,

Таблица 16

Зависимость между возрастом и отношением к моде

Отношение (степень согласия с утверждением) X	Возраст (Y)			$N(x_i)$
	молодые	среднего возраста	пожилые	
полное согласие	26	13	5	44
пожалуй, согласен	20	11	8	39
пожалуй, несогласен	9	10	20	39
полное несогласие	7	10	15	32
Всего	62	44	48	154

p=0,95, как это обычно принято в социальных исследованиях.

Далее. Каждую корреляционную таблицу можно охарактеризовать с помощью так называемого числа степеней свободы. Что это означает?

Нам заданы $N(x_i)$ и $N(y_j)$. Характер связи X с Y определит распределение объектов по $k \times l$ клеткам таблицы. Так как сумма частот клеток строки (как и столбца) фиксирована, то на распределение объектов по клеткам в каждой строке и в каждом столбце наложено по одному ограничению. Общее число ограничений $k+l$ должно быть уменьшено на 1, так как эти ограничения не независимы: сумма итогов столбцов равна сумме итогов строк (и равна N). Следовательно, на распределение объектов по $k \bullet l$ клеткам таблицы наложено $k+l-1$ ограничение. Величина $f = kl - (k + l - 1) = (k - 1)(l - 1)$ называется числом степеней свободы корреляционной таблицы.

Для разных p и f составлены специальные математические таблицы[6], по которым можно найти величину χ_0^2,

[73]

[5] О понятии вероятности см. Приложение 1.

[6] Приложение 3, таблица Б (χ_0^2).

обладающую таким свойством: для данной корреляционной таблицы (χ^2, f) с вероятностью p[7] можно утверждать, что отклонение теоретической таблицы от эмпирической носит случайный характер, если $\chi^2 \le \chi_0^2$. Если же $\chi^2 > \chi_0^2$, то расхождение нельзя считать случайным. Приведем пример вычисления χ^2.

Таблица 17

Пример расчета χ^2

Номер клетки	N_{ij}	N_{ij}^0	$N_{ij} - N_{ij}^0$	$(N_{ij} - N_{ij}^0)^2$	$\frac{(N_{ij} - N_{ij}^0)^2}{N_{ij}^0}$
1	26	17,7	8,3	68,89	3,89
2	13	12,7	0,3	0,09	0,01
3	5	13,6	-8,6	73,96	5,43
4	20	15,7	4.3	18,49	1.18
5	11	11,1	-0,1	0,01	0.00
6	8	12,2	-4,2	17,64	1.45
7	9	15,7	-6,7	44,89	2,86
8	10	11,1	-1,1	1,21	0,11
9	20	12,2	7,8	60,84	4.99
10	7	12,9	-5,9	34,81	2,70
11	10	9,1	0,9	0,81	0,09
12	15	10,0	5,0	25,0	2,50

Сумма цифр последней колонки — 25,21 — равна χ^2

Пример 13. Рассмотрим связь между признаками «отношение к моде» *(X)* и «возраст» *(Y).* Отношение будем измерять как степень согласия с утверждением: «Мода — это очень важно» (см. табл. 16), а возраст фиксировать в градациях: «молодые», «среднего возраста», «пожилые».

Рассмотрим эмпирическую корреляционную таблицу 17.

Составим расчетную таблицу для вычисления χ^2, нумеруя клетки корреляционной слева — направо, сверху — вниз.

f=3•2=6. Для p=0,95 χ_0^2=12,59; для p=0,99 χ_0^2=16,81. Следовательно, с p>0,99 можно утверждать,

[74]

[7] Часто при составлении таблиц вместо p используют величину q=1—p, которая называется *уровнем значимости.* Очевидно, p=0,95 соответствует уровень значимости 0,05 (т.е. 5%). В этом случае «в таблицу входят» по данному f и q=0,05 (5%). Именно этот уровень значимости чаще всего используется в социологии. В естественных науках обычно предпочтение отдают уровню 0,01 (1%).

что связь между отношением и возрастом есть. Установив статистический факт ее наличия, мы можем теперь обратиться к наполнению клеток таблицы, чтобы описать характер связи. Оказывается, что у молодых более позитивное отношение, у пожилых — более негативное.

Пример 14. При изучении связи между удовлетворенностью заработной платой (позиции шкалы: «удовлетворен», «трудно сказать», «не удовлетворен») и удовлетворенностью работой в целом (в тех же терминах) для молодых рабочих (возраст менее 30 лет) Одесского судоремонтного завода была получена следующая эмпирическая таблица 18.

Для нее $f = 2 \times 2 = 4$, χ^2 = 52,0 (проверьте!). Даже для p=0,99 χ_0^2=13,3, следовательно, гипотеза независимости признаков должна быть отвергнута с надежностью большей 0,99.

Вопрос о мере связи будет рассмотрен позднее.

Упражнение 25. Для рабочих в возрасте старше 30 лет аналогичная таблица имела вид (см. табл. 19).

Вычислить χ^2, найти χ_0^2 и сделать вывод о наличии или отсутствии связи между признаками. Ответ: связь есть, гипотеза независимости отвергается с p>0,99.

Итак, у молодых и пожилых работников есть связь между обсуждаемыми удовлетворенностями. Может возникнуть естественный вопрос: в каком случае связь большая? Чтобы ответить на него, нам придется рассмотреть ряд коэффициентов (Чупрова, Миркина, энтропийная мера связи — см. ниже), таким образом, мы еще несколько раз будем возвращаться к данным таблицы.

Упражнение 26. Показать, что в случае таблицы 2 × 2

$$\chi^2 = \frac{(N_{11} N_{22} - N_{21} N_{12})^2 N}{N(x_1) N(x_2) N(y_1) N(y_2)} \quad (\text{II},1,2)$$

Упражнение 27. Изучение распределения брачных пар по национальности мужа и жены в Казани[8] (1974 г.) дало таблицу 20.

Определить, есть ли связь между национальностью мужа и жены.

Вычислить χ^2 двумя способами: по общей формуле (III,1,1) и по (III,1,2). Ответ: 1052,6.

Так как f=(2—1)(2—1)=1, а для p=0,99 χ_0^2=6,63 намного меньше полученного значения, то с вероят-

[8] Рукавишников В.О. Население города. М., 1980, с.100.

Таблица 18

Связь между удовлетворенностью зарплатой *(X)* и удовлетворенностью работой *(Y)* для рабочих в возрасте до 30 лет

X	Y			
	y_1	y_2	y_3	$N(x_i)$
x_1	350	35	63	448
x_2	298	52	158	508
x_3	34	10	8	52
$N(y_j)$	682	97	229	1008

Таблица 19

Связь между удовлетворенностью зарплатой (*X*) и работой (*Y*) для рабочих в возрасте старше 30 лет

X	Y			
	y_1	y_2	y_3	$N(x_i)$
x_1	689	30	37	756
x_2	758	53	91	902
x_3	76	3	4	83
$N(y_j)$	1523	86	132	1741

Таблица 20

Связь между национальностями мужа и жены

Национальность жены	Национальность мужа		Всего
	русский	татарин	
Русская	924	51	975
Татарка	55	456	511
Всего	979	507	1486

[76]

ностью, большей чем 0,99, можно утверждать, что связь есть. О ее характере судят по распределению частот в клетках: семьи преимущественно гомогенны по национальности. Если бы семьи были преимущественно гетерогенны (например, если бы мы поменяли местами числа первой и второй строк таблицы), то χ^2 имел бы такое же высокое значение. Таким образом, χ^2 характеризует лишь степень тесноты связи, а не ее характер.

Таблица 21

Связь между квалификацией (*X*) и зарплатой (*Y*) у молодых рабочих

Квалификация (*X*)	Зарплата (*Y*), руб.						$N(x_i)$
	40-60	60-80	80-100	100-120	120-150	св. 150	
Низкая (x_1)	12	12	78	30	12	0	144
Средняя (x_2)	6	9	27	48	3	12	135
Высокая (x_3)	0	6	36	45	60	12	159
$N(y_j)$	18	27	141	123	105	24	438

Упражнение 28. Критерий χ^2 частот используется в социологическом исследовании «Человек и его работа»[9]. Приведем один из примеров. Изучался вопрос о связи между квалификацией x (x_1 — низкая, x_2— средняя, x_3 — высокая) и заработной платой y. Представляло интерес проверить, проявляется ли она в конкретном исследовании, осуществленном в Ленинграде (объект — молодые рабочие), так как общая закономерность отражает тенденцию, которая не исключает отклонений. Найти χ^2. Ответ: $\chi^2 = 92{,}2$

Для p=0,99 и f=2•5=10 χ_0^2=23,2<92,2. Следовательно, с p>0,99 можно утверждать, что расхождение эмпирических данных с гипотезой о независимости носит неслучайный характер, связь между признаками статистически подтверждается.

До сих пор речь шла о теоретических таблицах, построенных на основе гипотезы независимости, т.е. решался вопрос, есть ли связь между признаками. Однако теоретическая таблица может быть построена на основе предполагаемого характера распределения. Тогда с помощью χ^2 можно

[77]

[9] Человек и его работа. М., 1967, с. 352.

ответить на вопрос, соответствует ли эмпирическое распределение теоретическому:

$$\chi^2 = \sum_{i-1}^{n} \frac{(N_i - N_i^0)^2}{N_i^0} \quad \text{(II,1,3)}$$

где N_i *и* N_i^0 — эмпирическая и теоретическая частоты, а n — число вариантов. Формулу (II,1,1) можно рассматривать как частный случай формулы (II,1,3) для распределения с числом вариант $n=k \bullet l$. Теоретические частоты могут определяться на основании некоторой содержательной теории (в свое время таким способом была подтверждена справедливость корпускулярных законов наследственности: из теории определялось, каким должно быть соотношение сортов в опыте, а затем с помощью критерия χ^2 показывалось соответствие эмпирических частот теоретическим); на основании предположения о независимости (как было сделано ранее); из гипотезы о характере распределения (например, можно проверить соответствуют ли полученные данные предположению о нормальности распределения изучаемого признака). Так, в примере № 4 (рост 1000 мужчин) можно было бы найти средний рост, среднее квадратическое отклонение и по таблице нормального распределения определить, какая доля лиц должна попадать в каждый интервал при нормальном распределении. Умножая эту долю на число мужчин (1000) мы определили бы теоретические частоты, а затем, воспользовавшись формулой (II,1,3), можно было бы определить, отличается ли эмпирическое распределение от нормального.

Упражнение 29. В почтовом опросе работающего населения г. Киева было получено следующее распределение рабочих по разряду:

Частота	Разряд						Всего
	1	2	3	4	5	6	
Эмпирическая	19	83	145	171	219	153	790
Теоретическая	131,7	131,7	131,7	131,7	131,7	131,7	790

Проверим, может ли при таких данных действительное распределение (т.е. распределение для всех рабочих, а не только тех, кого мы опросили) быть равномерным? Если бы

[78]

распределение было бы равномерным, то рабочих каждого разряда было бы поровну, т.е. 790/6=131,7. Это и есть теоретические частоты. Отличается ли полученное распределение от равномерного? Ответ: χ^2=124,6 (отличается).

Критерий χ^2 дает возможность также сравнивать два ряда распределений и решать вопрос, случайно или нет различие между ними. При этом два распределения можно просто рассматривать как одну таблицу размера $2 \times k$ *(k* — число вариант). Рассмотрим этот вопрос на следующем примере.

Упражнение 30. При исследовании трудовых ресурсов Киева для экономии материальных и временных затрат нами была разработана следующая п роцедура[10]. На первом этапе мы провели репрезентативную для города по всем признакам анкеты выборку работающего населения, опросив около 900 респондентов методом интервью. Далее был проведен почтовый опрос, данные которого, как известно, подвержены различным смещениям. Чтобы устранить их, осуществлялся «ремонт» (коррекция) полученных в почтовом опросе 3,5 тысяч анкет по полу, возрасту и образованию, т.е. приведение всех пропорций по градациям этих признаков в соответствие с пропорциями в массиве, полученном путем интервью. Таким образом мы получили около 2,5 тыс. анкет «отремонтированного» массива. При этом возник вопрос, «отремонтировался» ли почтовый массив по остальным признакам, включенным в анкету, в частности, по признаку «тип рабочего места», (табл. 22).

Проверьте, отличаются ли эти два распределения. Чтобы ответить на этот вопрос требуется вычислить χ^2. Ответ: 2,84. Число степеней свободы равно 6. Проверить по таблице Б Приложения 3, что полученное расхождение незначимо, т.е. оно объясняется «игрой случая».

Можно, однако, поступить и иначе. Нас интересуют не просто различия распределений между собой, а то, насколько почтовый массив отличается от массива интервью. Данные интервью выступают в этом случае эталоном, теоретическим распределением. Итак, имеем эмпирическое распределение (почтовый массив) и теоретическое распределение (массив интервью). Но здесь есть небольшая сложность: теоретическое распределение должно иметь ту же сумму частот, что

[79]

[10] *Паниотто В. И., Яковенко Ю. И.* Некоторые способы совершенствования почтового опроса. — Социологические исследования, 1981, № 3.

и эмпирическое. Массив интервью дает нам лишь необходимые соотношения, по которым мы вычислим теоретические частоты: $N_i^0 = v_i^0 N$, где N_i^0 — теоретическая частота, v_i^0 — доля *i*-го варианта в распределении массива интервью, N — численность респондентов в почтовом опросе (т.е. 2459).

Таблица 22

Распределения респондентов по типу рабочих мест, полученные путем интервью и почтового опроса

Массивы	Тип рабочего места по характеру труда						
	Физический труд				Умственный труд		
	Неквалифицированный	Низко квалифицированный	Средней квалификации	Высокой квалификации	Не требующий высшего и среднего образования	Требующий среднего специального образования	Требующий высшего образования
Интервью (901 чел.)	43	43	158	143	107	120	287
«Отремонтированный» почтовый (2459 чел.)	134	127	409	415	318	315	741

Таким образом, $N_1^0 = \frac{43}{901} \bullet 2459$, $N_3^0 = \frac{158}{901} \bullet 2459$ и т.д.

Получаем следующее теоретическое распределение (с округлением до целых): 117, 117, 431, 390, 292, 328, 783. Сумма их будет уже не 901, а приблизительно 2459. По формуле (II,1,3): χ^2=11,1. Эта величина больше, чем рассчитанная ранее, но меньше 12,459 — критического значения для шести степеней свободы (т.е. различие незначимо). Как видим, результат зависит от формулировки проверяемой гипотезы (вопросы проверки гипотез подробнее будут рассмотрены в гл. V).

2. Коэффициенты, связанные с χ^2 (таблицы *k ➔ l)*

Прежде чем перейти к коэффициентам, базирующимся на критерии χ^2 Пирсона, приведем соотношение, которое понадобится нам в дальнейшем. Если учесть, что по опре-

делению $\sum_i \sum_j N_{ij} = \sum_i \sum_j N_{ij}^0 = N$, то из (II,1,1), возводя в квадрат числитель и расписывая выражение на три суммы, получаем:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{N_{ij}^2}{N_{ij}^0} - N \qquad \text{(II,2,1)}$$

Если связь функциональная (т.е. каждому *x* соответствует одно вполне определенное значение *y*), то без ограничения общности можно считать, что корреляционная таблица должна иметь диагональный вид. Пусть для определенности *k<l,* тогда $N_{ij} = \left\{ \begin{matrix} 0, i \neq j \\ N_{ij}, i = j = \overline{1,k} \end{matrix} \right\}$ и так как $N(x_i)=N(y_j)$, то $N_{ij}^0 = N_{ij}^2 / N$. Теперь просто найти $\chi^2_{\max}$. Подставляя N_{ij}^0 в (II,2,1) получаем: $\chi^2_{max} = N(k-1)$. При *k>l* аналогично $\chi^2_{max} = N(l-1)$. Таким образом,

$$\chi^2_{\max} = N \bullet \min(k-1, l-1) \qquad \text{(II,2,2)}$$

где *min(k – 1, l – 1)* обозначает наименьшее из двух чисел: *(k – 1)* и *(l – 1)*. (Отсюда, кстати, очевидно и определение величины *max(k – 1, l – 1)*, которая будет использована в дальнейшем).

Как мы видели, χ^2 — мера различия между эмпирической и теоретической таблицами, приходящаяся на все *N* объектов наблюдения.

Мера различия, приходящаяся на одно наблюдение, называется средней квадратической сопряженностью и обозначается φ^2: $\varphi^2 = \frac{\chi^2}{N}$.

Как и χ^2, $0 \le \varphi_2 < \infty$; отсутствие верхней границы у φ^2 не вполне удобно для коэффициента, характеризующего связь между признаками: обычно предпочтение отдают коэффициентам, принимающим значения между 0 и 1 (либо -1 и 1).

Пирсон предложил рассматривать величину

$$C = \sqrt{\frac{\varphi^2}{1 + \varphi^2}}, \qquad \text{(П,2,3)}$$

которая получила название *коэффициента средней квадратической сопряженности Пирсона.*

Легко видеть, что *C*=0 в случае отсутствия связи. В самом деле, при этом χ^2=0, следовательно φ^2=0 и *C*=0. Чем больше связь между признаками, тем больше С.

[81]

Но максимальное значение C не достигает 1. Чтобы устранить этот недостаток, целесообразно перейти к $C' = \frac{C}{C_{max}}$, где C_{max} — значение C при функциональной связи. Из (II,2,2) следует, что

$$C_{max} = \sqrt{\frac{min(k-1, l-1)}{1 + min(k-1, l-1)}}$$

Если таблица диагональная *(k = l)*, то $C_{\max} = \sqrt{\frac{k-1}{k}}$.

Прежде чем рассмотреть пример расчета χ^2, перепишем (II,2,1) с учетом выражения N_{ij}^0 через маргиналы $\frac{1}{N} N(x_i) \times N(y_i)$ в виде:

$$\chi^2 = N \sum_{i=1}^{k} \sum_{j=1}^{l} \left(\frac{N_{ij}^2}{N(x_i) N(y_j)} - 1 \right) \qquad \text{(II,2,1a)}$$

Пример 15. Для таблицы 20 рассчитать χ^2. По формуле (II,2,1 а) получаем

$$\chi^2 = 1486 \left(\frac{924^2}{979 \cdot 975} + \frac{51^2}{507 \cdot 975} + \frac{55^2}{979 \cdot 511} + \frac{456^2}{507 \cdot 511} - 1 \right) = 1052,6$$

Как видим, даже для таблицы 2×2 эта формула удобнее, чем (II,1,1) и (II,1,2), так как не требует оперирования большими числами, ею целесообразно пользоваться в подавляющем большинстве случаев.

Пример 16. Для данных таблицы 18 примера 14 рассчитать C, C_{max}, C'. Так как χ^2=52, получаем:

C=0,221; $C_{max} = \sqrt{\frac{2}{2+1}} = 0,816$; C'=0,271.

Упражнение 31. По данным примера 13 рассчитать C, C_{max}, C'. Ответ: 0,375; 0,816; 0,460.

Как мы видели, коэффициент, введенный Пирсоном, не может достигать 1. В свое время Чупров, стремясь исправить этот недостаток, предложил другой коэффициент, базирующийся на χ^2:

$$T = \sqrt{\frac{\chi^2}{N \sqrt{(k-1)(l-1)}}} \qquad \text{(П.2,4)}$$

Коэффициент Чупрова достигает максимального значения +1 в случае полной связи, но только при *k=l.*

[82]

Упражнение 32, Рассчитать T для полной связи при $k=l$. Указание: использовать (II,2,2).

Упражнение 33. По данным примера 14 вычислить коэффициент Чупрова для признаков удовлетворенность работой и удовлетворенность заработной платой (молодые рабочие). Заметим, что так как таблица квадратная, использование T вполне корректно. Ответ: 0,160.

Упражнение 34. То же для таблицы 19 (рабочие старших возрастных групп). Ответ: 0,078.

Сопоставим результаты двух последних упражнений. Как было ранее установлено, в обоих случаях связь между признаками есть, но можно ли сказать, в каком случае она больше? По-видимому, да: у молодых работников T больше, чем у работников более старших возрастных групп. Справедливость этого предварительного вывода в дальнейшем будет «подкреплена» с помощью различных других показателей.

Продолжим рассмотрение T. При $k\neq l$ $T_{max}<1$. Этот недостаток можно преодолеть так же, как и в случае C. Введем, следуя Крамеру, коэффициент $T_c = \frac{T}{T_{max}}$. Чтобы найти явное выражение T_c, вычислим T_{max}. Для этого воспользуемся (II,2,2) с учетом того, что $(k-1)(l-1)=min(k-1, l-1)\ max(k-1, l-1)$. Теперь (II,2,4) после простых преобразований дает:

$$T_{max} = \sqrt[4]{\frac{min(k-1, l-1)}{max(k-1, l-1)}};$$

$$T_c = T \cdot \sqrt[4]{\frac{max(k-1, l-1)}{min(k-1, l-1)}}$$

(Обратим внимание, что при выводе формулы для T_{max} и T_c, в изданном у нас переводе книги М. Кендалла и А. Стьюарта[11] допущена неточность: в обеих формулах приведен корень второй, а не четвертой степени).

Упражнение 35. По данным таблицы 22 рассчитать T и T_c. Ответ: 0,019; 0,029. $T_c \geq T$, причем равенство достигается при $k=l$. Коэффициент T_c называют коэффициентом Крамера, или обобщенным коэффициентом Чупрова. T_c существенно отличается от T для «вытянутых» таблиц.

Об использовании этих коэффициентов для факторного анализа связей между признаками и сопоставлении результатов, полученных при применении T и T_c, см. главу VI.

[83]

[11] *Кендалл М., Стьюарт А.* Статистические выводы и связи. М., 1973, с. 747.

Значения χ^2 и, следовательно, всех производных коэффициентов (φ^2, *C, T)* не чувствительны к последовательности значений x_i и y_j. Это дает возможность применять указанные меры даже для классификационных признаков, т.е. при самом слабом уровне измерения.

Для того чтобы выводы, получаемые при использовании обсуждаемых мер, были надежны, необходимо выполнение ряда условий. Как отмечают Дж.Юл и М.Кендалл[12], теоретические частоты N_{ij}^0 не должны быть меньше определенного минимума, в качестве которого они рекомендуют принять 10, полагая, что «предельный минимум» равен 5. Если в некоторых клетках теоретические частоты меньше, чем 5, нужно произвести объединение строк или столбцов. Общее число наблюдений *N* должно быть достаточно большим. Хотя трудно точно назвать его минимум, обычно доверяют результатам, если *N* не меньше 100 (конечно, если, скажем, *k*=5, а *l*=4, следовательно, число клеток 20, то *N* должно быть примерно равным 200, чтобы $N_{ij}^0 \geq 10$).

Значимость *C* и *T* определяется по значимости χ^2: если значим χ^2, то значимы и производные коэффициенты.

3. Таблицы 2 × 2. Коэффициенты ассоциации и контингенции, их связь с коэффициентами для таблиц *k* × *l*

Продолжим изучение коэффициентов, основанных на принципе совместного появления событий, обратившись к более простым ситуациям, чем раньше. Это позволит, в частности, лучше понять предыдущий материал, уяснить качественную основу его. Кроме того, мы изучим связи между новыми и уже рассмотренными коэффициентами. И, наконец, последующее изложение будет своеобразной «передышкой» для читателя, впервые столкнувшегося с изучением статистического материала. (Такому читателю будет полезно после изучения этого параграфа вернуться к предыдущим).

Оба коэффициента, о которых будет идти речь, применимы лишь к таблицам 2×2, т.е. в случае, когда данные сгруппированы дихотомически (табл. 23).

Напомним, что N_{12}, например, число индивидов, у которых $X=x_1$ и $Y=y_2$, $N(y_2)$ — число индивидов с $Y=y_2$ и любым *X*, а *N* — объем изучаемой совокупности.

[84]

[12] *Юл Дж., Кендалл М.* Теория статистики. М., I960, с. 526.

Для того чтобы перейти к рассмотрению связи, начнем с примера. Допустим, что нужно изучить связь между удовлетворенностью профессией —*Y* (y_1 — удовлетворен, y_2 — не удовлетворен) и фактической производительностью труда *X* (x_1 — высокая, x_2 — низкая). Часто приходится слышать утверждения типа: «Если удовлетворен профессией, то и производительность высокая». К таким посылкам и выводам обычно не придираются, считая их очевидными, не требую-

Таблица 23

Общий вид таблицы 2 × 2

X	*Y*		$N(x_i)$
	y_1	y_2	
x_1	N_{11}	N_{12}	$N(x_1)$
x_2	N_{21}	N_{22}	$N(x_2)$
$N(y_j)$	$N(y_1)$	$N(y_2)$	N

щими доказательства. Однако с подобными суждениями нельзя согласиться.

Как отмечалось, социальные явления многофакторны, а реальные связи далеки от тривиальности. Высокая производительность труда может соответствовать и высокой, и низкой удовлетворительности профессией (и наоборот). Речь идет пока об индивидуальных фактах. Что же касается статистических, изучением которых и занимается социолог, то здесь результат существенно определяется конкретной ситуацией, совокупностью многих условий жизнедеятельности. На разных совокупностях связь может быть разной — истина всегда конкретна. Заметим, что любой результат можно легко «объяснить», схватившись за один (подходящий) из множества влияющих факторов. Именно так легкомысленно поступают те, кто, узнав результат, говорят: «Это и так ясно, что тут исследовать?». Очевидно, необходимо уметь отличать общие рассуждения (и догадки!) от научно установленных фактов, даже если они относительно легко интерпретируются. Только такое знание может стать основой научных выводов, тем более — практических рекомендаций.

Пусть N=100 и 50 человек удовлетворены, а 50 — не удовлетворены профессией, у 20 — высокая, а у 80 — низкая производительность труда, т.е. корреляционная таблица

[85]

имеет вид (приведены только суммы частот, т.е. маргиналы):

X	Y		$N(x_i)$
	y_1	y_2	
x_1			20
x_2			80
$N(y_j)$	50	50	100

Пока мы знаем лишь маргиналы и не знаем, как распределены индивиды по клеткам таблицы, ничего нельзя сказать о связи. Информацию о ней несут только внутриклеточные частоты: лишь тогда, когда нам известны частоты *совместного появления* признаков, можно судить о связи.

Таблица 24

Зависимость между производительностью труда и удовлетворенностью профессией

Производительность труда (X)	Удовлетворенность профессией (Y)		$N(x_i)$
	удовлетворены y_1	не удовлетворены y_2	
Высокая — (x_1)	20	0	20
Низкая — (x_2)	30	50	80
$N(y_j)$	50	50	100

Следовательно, коэффициент, характеризующий ее, должен конструироваться из этих частот. Юл предложил описывать связь с помощью величины

$$Q = \frac{N_{11} N_{22} - N_{12} N_{21}}{N_{11} N_{22} + N_{12} N_{21}}$$

Прежде чем вычислить Q[13] и анализировать значения, принимаемые этим коэффициентом, рассмотрим содержательно несколько конкретных таблиц (табл. 24).

[86]

[13] Обозначение предложено Дж. Юлом в честь А. Кетле, одного из создателей научной статистики, впервые применившего количественные методы к изучению социальных явлений в своем — по оценке К.Маркса — «превосходном научном труде» «О человеке и развитии его способностей или опыт социальной физики», опубликованном в 1835 г. в Париже *(Маркс К., Энгельс Ф.* Соч., т. 8, с. 531).

В данной группе из 100 человек все, у кого высокая производительность труда, удовлетворены профессией (но не наоборот! об этом, впрочем, позднее), т.е. имеется полная определенность относительно удовлетворенности профессией у всех работников с высокой производительностью труда. Легко видеть, что при этом Q=1.

Далее будем рассматривать другие группы, для которых корреляционные таблицы имеют те же маргиналы, поэтому воспроизводить будем лишь внутриклеточные частоты.

19	1
31	49

а

15	5
35	45

б

10	10
40	40

в

Например, для таблицы ***а*** связь, очевидно, меньше, меньшим оказывается и Q=0,94. Для таблицы ***б*** связь еще меньше, и Q=+0,59. А для таблицы ***в*** связи между признаками нет: и у работников с высокой, и у работников с низкой производительностью труда числа удовлетворенных и неудовлетворенных профессией одинаковы. Соответственно обращается в нуль и Q.

Для того чтобы |Q| был равен 1, достаточно, чтобы одна из внутриклеточных частот обратилась в нуль. Например, при N_{12}=0 |Q|=1. Это значит, что если производительность высокая, то обязательно удовлетворен (разумеется, речь идет сданной гипотетической группе) профессией. Обратное неверно: если удовлетворен, то производительность может быть и высокая и низкая. Следовательно, Q — показатель односторонней связи. Если между значениями признаков

Таблица 25

Зависимость между учебой и участием в рационализации

Занятие учебой	Участие в рационализации		Всего
	Участвуют	Не участвуют	
учатся	29	93	122
не учатся	5	93	98
	34	186	120

[87]

допустимо упорядочение, как в нашем примере, то $Q>0$ соответствует прямой (высокой производительности отвечает высокая удовлетворенность), а $Q<0$ — обратной связи.

Упражнение 36. Вычислить Q для таблицы 25 (Ответ: Q=0,71). Связь есть. Она односторонняя: учеба влияет на участие в рационализации. Это же подтверждает значение Φ (см. ниже).

Коэффициент контингенции Ф по определению:

$$\Phi = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N(x_1)N(x_2)N(y_1)N(y_2)}}$$

В отличие от Q, который обращается в ±1, когда хотя бы одна внутриклеточная частота равна нулю, обращается в +1, когда $N_{12}=N_{2l}=0$, т.е. если — в нашем примере — все удовлетворенные профессией имеют высокую производительность, а неудовлетворенные — низкую (и наоборот!). Таким образом, Φ является показателем двусторонней связи. Соответственно: $|\Phi| \le |Q|$. Если $|\Phi| \ge 0{,}5$, то считают, что надежно установлена двусторонняя связь[14]. Если низкое значение $|Q|$ отвечает отсутствию связи *(*$|Q_{max}|=1$*)*, то низкое значение $|\Phi|$ может быть следствием маргинального эффекта: $|\Phi_{max}|$ часто меньше 1 (в этом можно убедиться на примерах). У разных таблиц разные Φ_{max}, поэтому Φ, рассчитанные для них, часто несопоставимы.

Можно показать, что нормировка Φ (переход к $\Phi' = \frac{\Phi}{\Phi_{max}}$) была бы незаконным усилением показателя связи. Если Φ мал, вычисляют Q, чтобы установить, есть ли хотя бы односторонняя связь. Так, для таблицы 25 Φ =0,26, а Q=0,71. Можно считать надежно установленной одностороннюю связь. (Вычисление этих коэффициентов составляет содержание *упражнения 37).*

Приведем примеры применения Q и Φ в социальных исследованиях (так как вычисления коэффициентов приводиться не будут, каждый из разбираемых примеров можно рассматривать как часть *упражнения 38*). Пусть X — место проживания, x_1 — город, x_2 — сельская местность, а Y — уровень образования, y_1 — высшее, среднее (оконченное и неоконченное), y_2 — начальное (оконченное и неоконченное). В таблицах, которые мы приведем по книге Ф.М. Бородкина «Статистическая оценка связей между экономиче-

[14] Более строго значимость Φ и Q определяют с помощью критерия χ^2.

скими показателями» (М., 1968), количества выражены в миллионах человек.

Итак, по данным на 1939 г.:

X	Y		$N(x_i)$
	y_1	y_2	
x_1	10,76	45,34	56,10
x_2	5,10	109,40	114,50
$N(y_j)$	15,86	154,74	170,60

Распределение маргиналов сходно, Φ=0,24; Q=0,67 Связь есть, она существенная, односторонняя (если житель сельской местности, то в большинстве случаев — человек с низким образовательным уровнем).

По данным за 1959 г.:

X	Y		$N(x_i)$
	y_1	y_2	
x_1	37,63	62,17	99,80
x_2	21,08	87,92	109,00
$N(y_j)$	58,71	150,09	208,80

Теперь Q=0,43. Это меньше, чем Q для предыдущей таблицы (1939). Следовательно, как видим, различия в уровне образования с течением времени стираются, хотя и остаются.

По данным 1939 г. и 1959 г. проследим связь между обсуждаемыми признаками у мужчин и у женщин в отдельности.

Для мужчин соответствующая таблица (1939 г.):

X	Y		$N(x_i)$
	y_1	y_2	
x_1	5,58	23,32	28,90
x_2	3,27	59,23	62,50
$N(y_j)$	8,85	82,55	91,40
Q=0,63			

[89]

Для женщин:

X	Y		$N(x_i)$
	y_1	y_2	
x_1	5,18	26,31	31,49
x_2	1,83	65,95	67,78
$N(y_j)$	7,01	92,26	99,27
Q=0,75			

Таким образом, различие в уровне образования горожанок и сельских жительниц более существенное, чем у мужчин.

Проследим динамику. Из соответствующих таблиц (данные 1959 г.) для мужчин Q=0,38, для женщин Q=0,47. Сделанный ранее вывод сохраняется, но связь становится менее существенной: и у мужчин, и у женщин с течением времени стираются различия образовательного уровня горожан и сельских жителей, хотя у женщин эти различия остаются несколько большими.

А теперь обратимся к материалам переписи 1970 г. В III томе «Итогов всесоюзной переписи населения 1970 года» — «Уровень образования населения СССР» (Москва, 1972, с. 206) — приводятся такие данные: на 1000 человек городского населения приходится 592 чел. с образованием выше начального, на 1000 же человек сельского населения — 332. Очевидно, по этим данным нельзя непосредственно рассчитать Q, так как численность городского и сельского населения неодинакова.

По данным V тома «Переписи» в городах проживало 135,33, а в селах—106,11 миллионов человек. Нужно, очевидно, 135,33 разделить в отношении 592:408, а 106,11 — в отношении 332:668. В результате получаем таблицу:

X	Y		$N(x_i)$
	y_1	y_2	
x_1	80,12	55,21	135,33
x_2	35,23	70,88	106,11
$N(y_j)$	115,35	126,09	241,44
Q=0,49			

[90]

Упражнение 39. Мужское население городов составляет 62,68 млн. чел., сельской местности — 48,50. На 1000 мужчин, проживающих в городе, приходится 621 чел. с образованием выше начального, а в сельской местности — 388 чел.

Составить таблицу, вычислить *Q*.

Ответ: *Q*=0,44.

Упражнение 40. Женское население городов составляет 72,65 млн. чел., сельское — 57,60 млн. чел. На 1000женщин, проживающих в городах, приходится 568 чел. с образованием выше начального, в сельской — 296. Составить таблицу, вычислить *Q*.

Ответ: *Q*=0,52.

Для контроля всех таблиц: все население СССР в 1970 г. составляло 241,44 млн. чел., в том числе: женщин — 130,26 млн. чел., мужчин — 111,18 млн. чел.

Рассмотрим полученные результаты. Грамотность населения СССР неуклонно возрастает, однако различие в уровне образования жителей городов и сельских местностей остаются: темпы роста образовательного уровня в городах выше.

Некоторое увеличение *Q* для таблиц 1970 г. по сравнению с *Q* для таблиц 1959 г. связано, по-видимому, с продолжающимся оттоком молодежи из сельских местностей в города. Из села уходят преимущественно молодые люди со средним (оконченным и неоконченным) образованием, в селе, таким образом, увеличивается доля тех, у кого образование не выше начального (это, в основном, старшие возрастные группы населения)[15].

Сделаем одно очень существенное замечание. Изучаемые социологами совокупности часто оказываются весьма разнородными. Например, рабочие предприятия —люди разных профессий, разного пола, возраста, образования и т.д. При достаточно разнородной совокупности могут возникать кажущиеся связи, либо оказаться скрытыми действительные. Поясним это примером.

Пример 17. Допустим, что некоторая совокупность может быть описана с помощью корреляционной таблицы такого вида:

[91]

[15] Любопытный пример применения Q в социологии читатель может найти в статье С. Железко «Факторы стабилизации кадров на строительстве БАМа» (Социологические исследования, 1980, № 1, с. 84—87).

X	Y		$N(x_i)$
	y_1	y_2	
x_1	300	300	600
x_2	200	200	400
$N(y_j)$	500	500	1000

Для нее Q, очевидно, равно нулю.

Предположим, что эта совокупность может быть по какому-либо признаку (например, по полу) разбита на 2 совокупности:

а

X	Y		$N(x_i)$
	y_1	y_2	
x_1	100	50	150
x_2	50	150	200
$N(y_j)$	150	200	350

б

X	Y		$N(x_i)$
	y_1	y_2	
x_1	200	250	450
x_2	150	50	200
$N(y_j)$	350	300	650

Для первой Q=+0,71, для второй Q= — 0,58.

Таким образом, для одной подсовокупности (например, для мужчин) связь между признаками X и Y положительная, а для другой (для женщин) — отрицательная.

Этот пример формально иллюстрирует случай, когда связь оказалась скрытой.

Несложно сконструировать пример, когда возникают кажущиеся связи. Дело здесь, конечно, не в «подгонке» соответствующих таблиц, а в том, что подобные эффекты могут *иметь место в реальной ситуации.* Как избежать их?

Детальные рекомендации давать трудно, но важно, чтобы социолог не применял коэффициенты бездумно. Нужно осмысливать изучаемую ситуацию, уделять большое внимание однородности изучаемых социальных общностей (это не означает, конечно, что нельзя выделять и исследовать параллельно разнородные группы).

И, наконец, о связях коэффициентов Q и Φ с φ и C.

С учетом (II,1,2) и (I,3,2) легко видеть, что для таблиц 2×2: $\Phi^2 = \frac{\chi^2}{N}$. С другой стороны, как мы видели,

для таблиц $k \times l$: $\varphi^2 = \frac{\chi^2}{N}$, т.е. φ^2 является обобщением Φ на случай корреляционных таблиц общего вида.

В качестве своеобразного обобщения Q и Φ можно рассматривать и коэффициент средней квадрата ческой сопряженности C.

О связи Φ коэффициента с коэффициентом Кендэла см. в конце §6 этой главы.

4. Коэффициент ранговой корреляции Спирмена

Рассмотренные ранее меры базируются, как отмечалось, на принципе *совместного* появления событий. Они пригодны для любых признаков — метрических, порядковых и даже номинальных.

Для метрических и порядковых признаков могут использоваться меры, основанные на принципе *ковариации.* Говорят, что переменные ковариантны, если вариации одной соответствует вариациям другой. Принцип ковариации, другими словами, основан на изучении совместных изменений в значениях признаков. Ясно, что его можно использовать для количественных данных, однако социальные признаки зачастую допускают только упорядочение. Например, ориентации, оценки, удовлетворенности, являющиеся собственно социологическими переменными, по существу измеряются с помощью шкал порядка: соответствующие эмпирические процедуры, как мы видели, дают возможность сказать, что индивид A более удовлетворен, чем B, своей специальностью, например, но не позволяют сказать на сколько (тем более — во сколько раз) больше.

Если совокупность упорядочена по двум (или более) признакам и изменению одного признака соответствует изменение другого, то говорят о наличии корреляции между ними. Чем можно измерить эту корреляцию?

Спирменовский коэффициент корреляции рангов. Допустим, что N индивидов могут быть упорядочены как по признаку X, так и по признаку Y. Пусть $R_i^{(x)}$ — ранг i-го индивида по признаку $X(i = \overline{1,N})$, а $R_i^{(y)}$ — по Y. Мерой несовпадения их является величина $d_i = R_i^{(x)} - R_i^{(y)}$. Во избежание эффекта компенсации, как и ранее, при переходе к полной мере возведем d_i, в квадрат и сложим, т.е. рассмотрим $\sum_{i=1}^{N} d_i^2$.

[93]

Потребуем далее, чтобы: 1) искомый коэффициент корреляции рангов обращался в +1, если все ранги совпадают, и 2) в (—1), если ранговые ряды имеют обратное направление (так, для N=5, $R_i^{(x)}$=1, 2, 3, 4, 5, а $R_i^{(y)}$=5, 4, 3, 2, 1).

Станем искать этот коэффициент в виде $1 - f\sum_{i=1}^{N} d_i^2$

(величину f мы найдем чуть позднее), тогда первое требование выполняется автоматически: если ранговые ряды идентичны, то $\sum_{i=1}^{N} d_i^2 = 0$

и коэффициент равен 1. Выберем f так, чтобы удовлетворить второму требованию.

Допустим, сперва, что N четно. Например, для N=6 имеем:

$R_i^{(x)}$	1	2	3	4	5	6
$R_i^{(y)}$	6	5	4	3	2	1
d_i^2	5^2	3^2	1^2	1^2	3^2	5^2

При N=2k:

$R_i^{(x)}$	1	2	…	k–1	k	k+1	k+2	…	2k–1	2k
$R_i^{(y)}$	2k	2k–1	…	k+2	k+1	k	k–1	…	2	1
d_i^2	$(2k-1)^2$	$(2k-3)^2$	…	3^2	1^2	1^2	3^2	…	$(2k-3)^2$	$(2k-1)^2$

$$\sum d_i^2 = 2\left[1^2 + 3^2 + ... + (2k-1)^2\right] = \frac{1}{3}k(4k^2-1)$$

(см. Приложение 2), следовательно, $\sum d_i^2 = \frac{1}{6}N(N^2-1)$

Упражнение 41. Вычислить $\sum d_i^2$ при N=2k+1

Указание: сперва рассмотреть N=7, по аналогии с N=6 (см. выше), а затем N=2k+1; воспользоваться соотношением

$$1^2 + 2^2 + 3^2 + ... + k^2 = \frac{k(k+1)(2k+1)}{6}$$

(см. Приложение № 2). Ответ: $\frac{N(N^2-1)}{6}$.

Если положить $f = \frac{6}{N(N^2-1)}$, то коэффициент

$$\rho = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)}$$

будет обладать требуемым свойством.

[94]

Пример 18. Изучая связь между субъективным отношением работников к труду (удовлетворенность работой) и объективным (текучесть), мы, в частности, оценивали ее с помощью коэффициента Спирмена в «сечении» возраст. Во второй колонке таблицы 26 значения индексов удовлетворенности работой различных возрастных групп работников Одесского судоремонтного завода (ОСРЗ).

Кроме того, нами изучалась текучесть работников. Каждая возрастная группа характеризуется определенным коэф-

Таблица 26

Вычисление коэффициента Спирмена ρ

Возрастные группы	Индексы удовлетворенности работой i_p	Коэффициент текучести K_T, %	Ранги по X (i_p)	Ранги по Y (K_T)	$\|d\|$	d^2
1	2	3	4	5	6	7
до 18 лет	0,57	12,9	5	5	0	0
18 – 19	0,38	13,0	7	4	3	9
20 – 21	0,35	17,1	8	3	5	25
22 – 24	0,24	37,1	9	1	8	64
25 – 30	0,39	19,9	6	2	4	16
31 – 40	0,59	7,9	4	6	2	4
41 – 50	0,69	5,6	3	9	6	36
51 – 60	0,76	6,1	2	8	6	36
свыше 60 лет	0,77	6,4	1	7	6	36

фициентом текучести, значения этих коэффициентов находятся в третьей колонке

$$\rho = 1 - \frac{6 \bullet 266}{8 \bullet 9 \bullet 10} = -0{,}88$$

Упражнение 42. В «сечении» стаж были получены такие данные:

Стаж, лет	i_p	K_T, %
До 5	0,41	26,5
5-10	0,46	15,1
10-15	0,58	3,6
15-20	0,65	3,3
Свыше 20	0,73	1,3

Вычислить ρ.

Аналогичный результат был получен в «сечении» образовательных групп. Все это позволило заключить, что между выделенными признаками имеется обратная (отрицательная)

[95]

связь, т.е. субъективное и объективное отношение к труду тесно связаны.

До сих пор предполагалось, что все ранги различны. Может, однако, случиться, что с точностью нашего измерения ранги у нескольких индивидов окажутся одинаковыми. Если, например, данный признак в максимальной степени присущ *A* и *B,* то каждому мы присвоим ранг 1,5=(1+2)/2.

Если, например, вслед за ними идут C, D, E с одинаковой степенью признака, то каждому из индивидов мы присвоим ранг (3+4+5)/3=4. В таких случаях говорят об объединении рангов. Выведенная формула для случая объединенных рангов может быть обобщена (мы это сделаем в §5). Сейчас же укажем конечный результат. Если среди рангов по *X* встречается *p* различных объединений и в *s*-ом объединено t_s объектов (рангов), где $s = \overline{1, p}$, а среди рангов Y имеется *q* объединений по u_r объектов в каждом, где $r = \overline{1, q}$,

то $\rho = \dfrac{\frac{N^3-N}{6} - \sum d_i^2 - T_x - T_y}{\sqrt{(\frac{N^3-N}{6} - 2T_x)(\frac{N^3-N}{6} - 2T_y)}}$,

где

$$T_x = \frac{1}{12}\sum_{s=1}^{p} t_s(t_s^2 - 1); \; T_y = \frac{1}{12}\sum_{r=1}^{q} u_r(u_r^2 - 1).$$

Последняя формула, как легко видеть, в случае отсутствия объединений легко превращается в ранее полученную (11,4,1). Рассмотрение ранговой корреляции на этом мы не заканчиваем. В дальнейшем (§ 6) будет введен другой коэффициент ранговой корреляции, предложенный Кендэлом.

Кроме того, для уяснения смысла коэффициента Спирмена мы проследим его связь с так называемым коэффициентом парной корреляции Пирсона — Браве. Это позволит уточнить условия и область применения спирменовского коэффициента.

Коэффициент Пирсона — Браве, к рассмотрению которого мы переходим, также основан на принципе ковариации. Он применим только к количественным признакам.

[96]

5. Коэффициент парной корреляции и его связь с другими коэффициентами

Вначале придем к коэффициенту парной корреляции полукачественным образом (аналогично выводу ρ). Такой нестрогий вывод, однако, полезен, так как помогает понять смысл коэффициента.

Итак, данный коэффициент один из показателей корреляционной связи. Основные задачи корреляционного анализа состоят в установлении формы связи, т.е. определении вида корреляционного уравнения (как это делается, мы рассмотрим в следующей главе), а также в определении тесноты, «силы» связи, т.е. оценке степени рассеяния эмпирических значений *y* около линии регрессии для разных *x*.

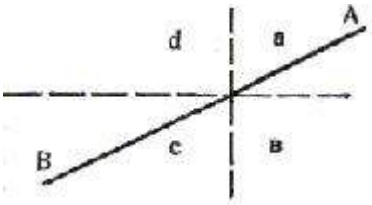


Рис. 20. Области корреляционного поля

Мерой тесноты связи в случае линейной корреляционной зависимости, как мы увидим, является коэффициент парной корреляции, а при криволинейной зависимости — корреляционное отношение.

Остановимся несколько подробнее на понятии тесноты связи. Если нанести все пары *x* и *y* в виде точек на плоскости, то получится, как упоминалось, корреляционное поле. Его точки располагаются в окрестности линии регрессии, компактно или разбросано. Поясним это примером.

Допустим, что сопоставляется возраст учащегося *(Y)* и год обучения *(X)*. Если речь идет о школьниках, то зависимость прямая функциональная: так в первом классе, в основном, дети семилетнего возраста, во втором — восьмилетнего и т.д. Второгодничество, обусловленное болезнями, реже — плохой успеваемостью, несколько «размывает» зависимость, делает ее корреляционной, но точки корреляционного поля тесно располагаются в окрестности прямой регрессии.

Перейдем к рассмотрению обучения в вузе. Не все студенты—вчерашние школьники, многие приходят в вуз после армии, работы в народном хозяйстве, поэтому разброс значений возраста студентов на разных курсах значительно больше, чем в разных классах школы: корреляционное поле «размывается».

[97]

Процент приходящих в аспирантуру после работы значительно выше, причем приходят люди после разных перерывов в учебе, разброс значений возраста аспирантов на каждом курсе выше, чем у студентов, корреляционное поле еще более «размыто».

Охарактеризовать «размытость» этого поля можно с помощью отклонений индивидуальных эмпирических значений от средних, т.е. $x_i - \bar{x}$ и $y_i - \bar{y}$. Если значению *x,* меньшему среднего, соответствует значение, *y* тоже меньшее среднего (а большему — большее), то это свидетельствует об упорядоченности, о наличии связи, мерой которой может служить величина

$$S = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}).$$

Действительно, чем больше совпадений знаков упомянутых отклонений, т.е. чем больше упорядоченность, тем больше *S*. При несовпадении знаков отклонений в сумме появляются отрицательные слагаемые, и она уменьшается. Если связи нет, то положительные и отрицательные слагаемые примерно уравновесятся и сумма *S* будет близка к нулю.

Заметим, что пока речь шла о положительной связи. Связь может быть отрицательной, в этом случае знаки отклонений Δx_i и Δy_i преимущественно совпадать не будут и величина *S* становится отрицательной. Теперь совпадения знаков индивидуальных отклонений уменьшают *S* по абсолютной величине, приближая ее к нулю.

Перейдем к графической интерпретации.

На рис. 20 прямые $y = \bar{y}$ и $x = \bar{x}$ разбивают координатную плоскость на 4 части: *a, b, c, d.* Положительность *S* означает преимущественное расположение точек корреляционного поля в областях *a* и *c* (отрицательность — *b* и *d).* Величина *S* близка к нулю, если поле равномерно «размазано».

Рассмотрим, для определенности, $S>0$. Чем больше *S*, тем более упорядочено корреляционное поле. В каком случае упорядоченность максимальна? Если зависимость функциональная, прямолинейная, то, очевидно, когда все точки лежат на прямой, скажем, *(AB).*

В качестве меры тесноты связи удобно рассматривать отношение *S* к его максимально возможному значению. Это отношение *r*, называемое коэффициентом парной корреляции, очевидно, принимает значение +1, если связь прямо-

линейная положительная; –1 — если прямолинейная отрицательная; 0 — если связи нет[16]. Таким образом, для того чтобы полностью определить r, остается найти максимальное значение величины S. Так как при прямолинейной связи $y_i=ax_i+b$, то $\Delta y_i=y_i-y=a\bullet\Delta x_i$, откуда $\Delta x_i=\frac{1}{a}\Delta y_i$ Поэтому $S_{max}=a\sum(\Delta x_i)^2$ и в то же время $S_{max}=\frac{1}{a}\sum(\Delta y_i)^2$. Чтобы освободиться от a, запишем $S_{max}=\sqrt{S_{max}\cdot S_{min}}=\sqrt{\sum(\Delta x_i)^2\sum(\Delta y_i)^2}$.

Окончательно:

$$r=\frac{\sum\Delta x_i\Delta y_i}{\sqrt{\sum(\Delta x_i)^2\cdot\sum(\Delta y_i)^2}} \qquad \text{(II,5,1)}$$

Формулу (II,5,1) можно записать в виде

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N\sigma_x\sigma_y}, \qquad \text{(П,5,2)}$$

если использовать определение среднего квадратического отклонения.

Упражнение 43. Показать, что коэффициент парной корреляции может быть представлен в виде

$$r=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sqrt{(\overline{x^2}-\bar{x}^2)(\overline{y^2}-\bar{y}^2)}}.$$

Указание: использовать соотношение (I,4,3) для обеих переменных — X и Y, а также определение средних.

Для уяснения смысла r полезно обратиться к некоторым частным случаям, где связь просматривается наглядно.

1. Пусть X и Y принимают такие значения:

X	1	2	3	4	5
Y	12	13	14	15	16

Ясно, что $y=x+11$, т.е. имеет место прямолинейная положительная связь

$N=5,\bar{x}=3,\bar{y}=14,\sigma_x=\sigma_y=\sqrt{2},\sum\Delta x\times\Delta y=10,r=1$ (Вычисление приведенных значений составляет содержание *упражнения 44*).

2. Пусть X и Y принимают значения:

X	*1*	*2*	*3*	*4*	*5*
Y	*16*	*15*	*14*	*13*	*12*

$r=-1$ *(Упражнение 45.* Показать это самостоятельно).

[99]

[16] Ниже мы остановимся подробнее на случае $S=0$.

Упражнение 46. Для данных следующей таблицы вычислить *r*.

X	*1*	*2*	*3*	*4*	*5*
Y	*13*	*16*	*14*	*12*	*15*

Ответ: *r*=0

Специально остановимся на рассмотрении случаев, когда *r* равно или близко к нулю. Всегда ли это означает отсутствие связи?

Вообще говоря, нет. Вспомним, что мера *r* приспособлена к изучению прямолинейных зависимостей, *r* может быть ма-

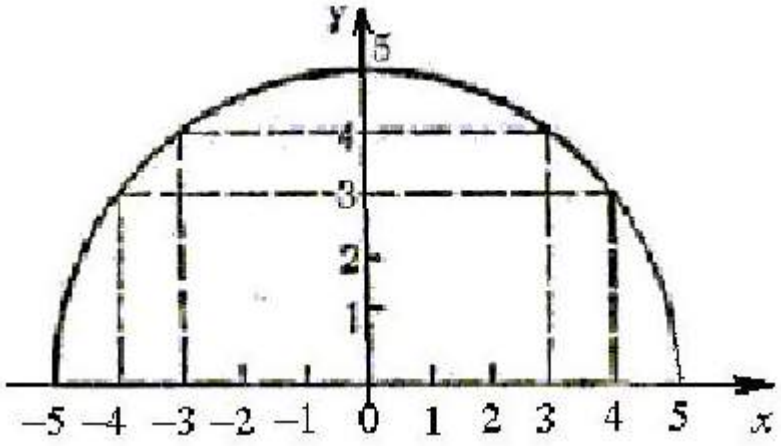


Рис.21. Криволинейная функциональная зависимость

лым или даже равным нулю не потому, что связи нет, а потому, что она криволинейна. Это помогает понять простой пример. Пусть X и Y заданы с помощью следующей таблицы:

X	*–5*	*–4*	*–3*	*0*	*3*	*4*	*5*
Y	*0*	*3*	*4*	*5*	*4*	*3*	*0*

Вычислим *r* для этих данных. Ясно, что $x=0; \overline{xy}=0,$ следовательно, по (II,5,3) *r*=0.

Одновременно легко видеть, что рассматриваемые величины *x* и *y* связаны функционально: $y=\sqrt{25-x^2}$.

Представим эту зависимость графически (рис. 21).

Таким образом, в нашем случае при *r*=0 имеет место криволинейная (даже функциональная) зависимость.

Итак, если *r*=0 (либо близко к нулю), то это означает отсутствие прямолинейной связи, но может иметь место криволинейная (обычно корреляционная) связь между изучаемыми величинами.

Упражнение 47. Показать, что в случае корреляционной таблицы $\{N_{ij}\}$ коэффициент корреляции Пирсона — Браве

[100]

принимает вид:

$$r=\frac{N\sum_{i=1}^{k}\sum_{j=1}^{l}N_{ij}x_iy_i-\sum_{i=1}^{k}N(x_i)x_i\cdot\sum_{j=1}^{l}N(y_j)y_j}{\sqrt{N\sum_{i=1}^{k}N(x_i)x_i^2-\left[\sum_{i=1}^{k}N(x_i)x_i\right]^2\times N\sum_{j=1}^{l}N(y_j)y_j^2-\left[\sum_{j=1}^{l}N(y_j)y_j\right]^2}} \quad \text{(II,5,4)}$$

Указание: использовать (II,5,3).

Как уже отмечалось, r=1 означает наличие положительной прямолинейной связи, r=—1 — отрицательной, а r=0 — отсутствие прямолинейной корреляционной связи. Значения, получаемые на практике, обычно таковы, что $0<|r|<1$. Вопрос о существенности r см. в § 8 главы V.

Заметим, что без обоснования линейности связи использование r не является законным, хотя и получило широкое распространение.

Для нелинейных зависимостей, какими часто являются социальные, нужно применять корреляционное отношение. Этот коэффициент будет подробно проанализирован в следующей главе. Здесь же мы придем к нему из качественных соображений. В случае корреляционной связи каждому x_i соответствует

$$\overline{y_i}=\frac{\sum_{j=1}^{l}N_{ij}y_j}{N(x_i)},$$

так называемое условное среднее (условие: X=x_i).

Вообще говоря, $\overline{y_i}$ не совпадают со средним значением

$$\overline{y}=\frac{1}{N}\sum_{j=1}^{l}N(y_j)y_j.$$

Мерой отклонения эмпирических $\overline{y_i}$ от $\overline{y}$ может служить величина

$$\sigma_{\overline{y}}=\sqrt{\frac{1}{N}\sum_{i=1}^{k}N(x_i)(\overline{y_i}-\overline{y})^2},$$

которая в терминах § 3 главы II может рассматриваться как межгрупповая дисперсия (там эта дисперсия обозначалась δ).

[101]

Корреляционным отношением η называется отношение $\sigma_{\bar{y}}$ и σ_y. Покажем, что эта величина действительно имеет смысл меры тесноты корреляции в случае криволинейной зависимости. Если зависимости нет, то $\overline{y_i}$ не будет отличаться от $\overline{y}$, т.е. $\sigma_{\bar{y}} = 0$ и η = 0.

Если зависимость функциональная, т.е. каждому *X* соответствует одно определенное значение *V,* то частные дисперсии $\sigma_i^2(y) = 0$ $(i = 1, K)$ и, следовательно, их средняя $\overline{\sigma^2}$ тоже равна 0.

Поэтому теорема сложения дисперсий (I,4,8) в этом случае дает: $\sigma_y^2 = \sigma_{\bar{y}}^2$, т.е. η = 1.

Итак, $0 \leq \eta \leq 1$, где 0 соответствует отсутствию связи, 1 – функциональной, а η, удовлетворяющие условию $0 < \eta < 1$, – корреляционной. Чем ближе η к 1, тем теснее связь, тем ближе она к функциональной.

Вернемся к рассмотрению r. Не является законным использование r также в случае, когда признаки не количественные. Рассмотрим один из типичных примеров. В исследовании «Человек и его работа», в частности, изучалась связь между такими признаками, как содержание труда и удовлетворенность специальностью. Профессии группировались по содержанию труда с учетом критериев, связанных с творческими возможностями трудовой деятельности (уровень механизации, уровень квалификации, соотношение затрат умственного и физического труда)[17]. Были выделены такие группы: 1) ручной труд, не требующий специальной подготовки; 2) труд на конвейере; 3) механизированный труд (станочный); 4) автоматчики без навыков наладки; 5) ручной труд, требующий высшей квалификации; 6) пультовикн-наладчики.

Ясно, что эти группы – пункты в лучшем случае порядковой шкалы. Удовлетворенность специальностью определялась по ответам на вопросы анкеты, упорядоченным по схеме «логического квадрата», следовательно, также по порядковой шкале. Корреляция же между выделенными признаками изучалась с помощью коэффициента Пирсона – Браве, применимого лишь в случае метрических шкал, так как он базируется на понятии отклонения от среднего, которое имеет смысл лишь тогда, когда числа несут информацию об «абсолютной» интенсивности свойства. Таким образом,

[102]

[17] Человек и его работа. М., 1967, с.30-38.

использовалась информация, которой фактически исследователи не располагали. Наконец, r применялся без обоснования линейности связи. Покажем, что коэффициент Спирмена является коэффициентом Пирсона – Браве, примененным к рангам. Ранг по *X,* как и ранги по У, принимают значение от 1 до *N.* Среднее значение ранга $\frac{1+N}{2}$, а отклонение i-го ранга от среднего i - $\frac{1+N}{2}$.

Теперь $\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 \to \sum_{i=1}^{N}\left(i-\frac{1+N}{2}\right)^2=\frac{N^3-N}{12}$ (II, 5, 5)

(см. Приложение 2).

Аналогично

$$\sum\left(y_i-\bar{y}\right)^2 \to \frac{N^3-N}{12}$$

В обозначениях предыдущего параграфа:

$$d_i=R_i^{(x)}-R_i^{(y)}=\left(R_i^{(x)}-\frac{1+N}{2}\right)-\left(R_i^{(y)}-\frac{1+N}{2}\right),$$

$$d_i^2=\left(R_i^{(x)}-\frac{1+N}{2}\right)^2+\left(R_i^{(y)}-\frac{1+N}{2}\right)^2-2\left(R_i^{(x)}-\frac{1+N}{2}\right)\left(R_i^{(y)}-\frac{1+N}{2}\right)$$

Отсюда

$$\sum_i\left(R_i^{(x)}-\frac{1+N}{2}\right)\left(R_i^{(y)}-\frac{1+N}{2}\right)=$$

$$=\frac{1}{2}\left[\sum_i\left(R_i^{(x)}-\frac{1+N}{2}\right)^2+\sum_i\left(R_i^{(y)}-\frac{1+N}{2}\right)^2-\sum_i d_i^2\right],$$

$$\sum_i\left(R_i^{(x)}-\frac{1+N}{2}\right)^2=\sum_i\left(i-\frac{1+N}{2}\right)^2=\frac{N^3-N}{12},$$

так как $R_i^{(x)}$ пробегает все значения от 1 до N.

Аналогично $\sum_i\left(R_i^{(y)}-\frac{1+N}{2}\right)^2=\frac{N^3-N}{12}$

[103]

следовательно, теперь

$$\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \to \frac{1}{2}\left(\frac{N^3 - N}{6} - \sum_i d_i^2\right)$$

Итак,

$$r \to 1 - \frac{6\sum d \frac{2}{i}}{N^3 - N} = \rho$$

Завершим здесь рассмотрение р выводом формулы для случая объединенных рангов.

У нас $i = \overline{1, N}$. Допустим, что ранги у нескольких объектов, например, с l+1 по l+t одинаковы. Каждому из этих t объектов естественно приписать средний ранг, который равен $l + \frac{1+t}{2}$. Найдем сумму квадратов объединенных рангов:

$$A = t\left(l + \frac{1+t}{2}\right)^2 = tl^2 + lt(t+1) + \frac{t(t+1)^2}{4}.$$

Если бы объединения не было, то сумма квадратов рангов тех же объектов была бы

$$B = (l+1)^2 + (l+2)^2 + ... + (l+t)^2 = tl^2 + lt(t+l) + \frac{t(t+l)(2t+1)}{6}.$$

Здесь мы воспользовались формулами Приложения 2.

Таким образом, при объединении рангов общая сумма квадратов окажется уменьшенной на величину $B - A = \frac{t(t^2-1)}{12}$.

Мы рассмотрели случай одного объединения (от $l+1$ до $l+t$), если объединений несколько, скажем, p, причем в s-ом случае объединено t_s рангов, то общее уменьшение

$$T_x = \sum_{w=1}^{p} \frac{t_s\left(t_s^2 - 1\right)}{12}, \text{ (II, 5, 6)}$$

если объединить ранги X.

Аналогичный вклад T_y дает объединение рангов по У:

$T_y = \sum_{r=1}^{q} \frac{u_r\left(u_r^2 - 1\right)}{12}$, где q – число объединений рангов Y, u_r - число рангов в r-ом объединении.

[104]

Введем эти поправки в формулу для р. Исходным при этом будет такое представление:

$$\rho = \frac{\sum_i \left(R_i^{(x)} - \frac{1+N}{2} \right)\left(R_i^{(y)} - \frac{1+N}{2} \right)}{\sqrt{\sum \left(R_i^{(x)} - \frac{1+N}{2} \right)^2 \sum_i \left(R_i^{(y)} - \frac{1+N}{2} \right)^2}}$$

Теперь

$$\sum_i \left(R_i^{(x)} - \frac{1+N}{2} \right)^2 \to \frac{N^3 - N}{12} - T_x;$$

$$\sum_i \left(R_i^{(y)} - \frac{1+N}{2} \right)^2 \to \frac{N^3 - N}{12} - T_y;$$

$$\sum_i \left(R_i^{(x)} - \frac{1+N}{2} \right)^2 \left(R_i^{(y)} - \frac{1+N}{2} \right) \to \frac{1}{2}\left[\frac{N^3 - N}{6} - \left(\sum d_i^2 + T_x + T_y \right) \right]$$

Таким образом,

$$\rho = \frac{\frac{N^3 - N}{6} - \sum_i d_i^2 - T_x - T_y}{\sqrt{\left(\frac{N^3 - N}{6} - 2T_x \right)\left(\frac{N^3 - N}{6} - 2T_y \right)}}. \qquad \text{(II, 5,7)}$$

В заключение параграфа приведем пример вычисления ρ с объединением рангов.

Пример 19. Изучая связь между положительными ответами на вопросы «интересная работа» (*X*) и «образование соответствует работе» (*Y*), социологи Казанского университета из 14 профессиональных групп рабочих (*N=14*) получили такие данные[18] (табл. 27, данные 1963г.):

$\sum {}^{2}_{di} = 286{,}5;$ р=0,354.

Имеем $T_x = 10{,}5$; $T_y = 1$;

Отметим, что в «Методике и технике…» и в «Статистических методах…», откуда взят этот пример, значение ρ равно 0,345. Полученное расхождение вызвано тем, что в обеих

[105]

[18] Методика и техника статистической обработки первичной социологической информации. М., 1968 г., с. 169, 170; этот же пример см.: Статистические методы анализа информации в социологических исследованиях. М., 1979, с.111, 112.

книгах использовалась следующая формула для расчета ρ:

$$\rho = 1 - \frac{6\left(\sum_i d_i^2 + T_z + T_y\right)}{N^3 - N} \qquad \text{(II, 5, 8)}$$

Как она соотносится с выведенной нами формулой? Преобразуя формулу (III,5,7), получаем:

$$\rho = \frac{\left(N^3 - N\right) - 6\left(\sum_i d_i^2 + T_x + T_y\right)}{\left(N^3 - N\right)\sqrt{\left(1 - \frac{12T_x}{N^3 - N}\right)\left(1 - \frac{12T_y}{N^3 - N}\right)}} \qquad \text{(II, 5, 9)}$$

Если в этой формуле пренебречь величинами, вычитающимися из 1 под корнем, то подкоренное выражение станет равно

Таблица 27

Пример вычисления коэффициента Спирмена ρ с объединением рангов

Номер группы	X (%)	Y(5)	R_i^X	R_i^Y	d_i	d_i^2
1	100	100	3	1	2	4
2	100	87,5	3	5,5	2,5	6,25
3	100	77	3	9	6	36
4	100	75	3	10	7	49
5	100	50	3	11,5	8,5	72,25
6	83,5	92	6,5	3	3,5	12,25
7	83,5	83	6,5	8	1,5	2,25
8	83,0	90	8	4	4,0	16,00
9	82,5	94,5	9	2	7,0	49,00
10	71,0	87,0	10	7	3,0	9,0
11	55,5	87,5	11	5,5	5,5	30,25
12	50,0	50,0	12	11,5	0,5	0,25
13	28,5	43,0	13	13	0	0
14	0	0	14	14	0	0

1 и (II,5,9) преобразуется в (II,5,8). Таким образом, (II,5,8) является приближенным выражением для (II,5,7).

Думается, что при наличии объединенных рангов ни (II,5,7), ни (II,5,8) не дают существенного упрощения расчетов, поэтому можно рекомендовать использовать для вычисления ρ формулу, по которой вычисляется r – (II,5,1), (II,5,2) или (II,5,3). Поскольку, как мы показали, ρ является коэффициентом r, примененным к рангам, результат будет тот же, что и по формуле (II,5,7). В частности, при использовании (II,5,1) для примера 19 получим 0,354.

[106]

6. Коэффициент ранговой корреляции Кендэла

В социологических исследованиях часто удается охарактеризовать объект не по абсолютной, а лишь по относительной интенсивности некоторого свойства (качественные признаки: оценки, удовлетворенность и т.д.). Таким образом, известна лишь последовательность, в которой располагаются объекты, т.е. каждый объект описывается с помощью рангов по каждому признаку. Ясно, что чем более согласованы ранговые ряды, тем больше связь между признаками.

Однако при строгом подходе ни *r*, ни ρ не могут использоваться как надежная мера связи двух качественных признаков (либо качественного и количественного), поскольку эмпирически не обоснованы отношения, используемые при построении этих коэффициентов.

Предложенный Кендэлом коэффициент строится на основе отношений типа «больше – меньше», справедливость которых установлена при построении шкал.

Рассмотрим логику вывода этого коэффициента. Пусть имеются N объектов. Из них можно выбрать $C_N^2 = \frac{N(N-1)}{2}$ различных пар. По предположению, известны ранги каждого объекта и по признаку *X* и по признаку Y.

Выделим пару объектов и сравним их ранги по одному признаку и по другому. Если по данному признаку ранги образуют прямой порядок (т.е. порядок натурального ряда), то паре приписывается +1, если обратный, то –1. Для выделенной пары соответствующие плюс – минус единицы (по признаку *X* и по признаку Y*)* перемножаются. Результат, очевидно, равен +1; если ранги пары обоих признаков расположены в одинаковой последовательности, и –1, если в обратной.

Если порядки рангов по обоим признакам у всех пар одинаковы, то сумма единиц, приписанных всем парам объектов, максимальна и равна числу пар. Если порядки рангов всех пар обратны, то $-C_N^2$. В общем случае $C_N^2 = P + Q$, где *P* – число положительных, *а Q.* – отрицательных единиц, приписанных парам при сопоставлении их рангов по обоим признакам.

Величина

$$\tau = \frac{P - Q}{\frac{1}{2} N(N-1)} \qquad \text{(II, 6,1)}$$

называется коэффициентом Кендэла.

[107]

Упражнение 48. 1. Убедиться, что в случае совпадения порядков рангов всех объектов по обоим признакам $\tau = +1$, а в случае обратного порядка $\tau = -1$.

2. Показать, что

а) $\tau = 1 - \dfrac{4Q}{N(N-1)}$ (II, 6, 2)

б) $\tau = \dfrac{4P}{N(N-1)} - 1$ (II, 6,3)

Из формулы (II, 6, 1) видно, что коэффициент τ представляет собой разность доли пар объектов, у которых совпадает порядок по обоим признакам (по отношению к числу всех пар) $\left(\dfrac{P}{\frac{1}{2}N(N-1)}\right)$ и доли пар объектов, у которых порядок не совпадает $\left(\dfrac{Q}{\frac{1}{2}N(N-1)}\right)$. Например, значение коэффициента 0,60 означает, что у 80% пар порядок объектов совпадает, а у 20% не совпадает (80% + 20% = 100%; 0,80 – 0,20 = 0,60). Т.е. τ можно трактовать как разность вероятностей совпадения и не совпадения порядков по обоим признакам для наугад выбранной пары объектов.

В общем случае расчет τ (точнее *P* или *Q*) даже для *N* порядка 10 оказывается громоздким. Покажем, как упростить вычисления.

Расположим объекты так, чтобы их ранги по *X* представили натуральный ряд. Так как оценки, приписываемые каждой паре этого ряда, положительные, значения «+1», входящие в *P,* будут порождаться только теми парами, ранги которых по *Y* образуют прямой порядок. Их легко подсчитать, сопоставляя последовательно ранги каждого объекта в ряду *Y* с остальными.

Покажем, как вычислять τ. Рассмотрим таблицу для $N = 10$:

Объекты	A	B	C	D	E	F	G	H	K	L
Ранг по X	6	4	2	10	9	3	1	5	7	8
Ранг по Y	8	7	6	10	5	2	1	3	4	9

Упорядочим ранги по *X:*

Объекты	G	C	F	B	H	A	K	L	E	D
Ранг по X	1	2	3	4	5	6	7	8	9	10
Ранг по Y	1	6	2	7	3	8	4	9	5	10

В ряду *Y* справа от 1 расположено 9 рангов, превосходящих 1, следовательно, 1 породит в *P* слагаемое 9. Справа от

[108]

6 стоят 4 ранга, превосходящих 6 (это 7, 8, 9, 10), т.е. в P войдет 4 и т.д. В итоге P=9+4+7+3+5+2+3+1+1 = 35 и с использованием (III,6,3) имеем: τ = + 0,56.

Упражнение 49. 12 объектов характеризуются двумя признаками X и Y. После упорядочения рангов по X таблица приняла следующий вид:

Ранг по X	1	2	3	4	5	6	7	8	9	10	11	12
Ранг по Y	3	4	1	5	2	11	9	6	7	8	10	12

Вычислить коэффициент Кендэла.

Для контроля вычислений: P = 53 (Q=13), τ =-0,24

Упражнение 50. Вычислить τ для признаков X и Y по следующим распределениям рангов:

Объекты	A	B	C	D	E	F	G	H	K	L
X–ранг	1	2	3	4	5	6	7	8	9	10
Y–ранг	7	10	4	1	6	8	9	5	2	3

Ответ: $\tau = -0,24$

Пример 20. При изучении связи между удовлетворенностью работой (J_p) и текучестью (K_T) работников в «сечении» возрастных групп были получены следующие результаты (ОСРЗ):

Возрастная группа	K_T (%)	J_p	ранг по X(K_T)	ранг по Y(J_p)
до 18 лет	12,9	0,57	5	5
18–19	13,0	0,38	4	7
20–21	17,1	0,35	3	8
22–24	37,1	0,24	1	9
25–30	19,9	0,39	2	6
31–40	7,9	0,59	6	4
41–50	5,6	0,69	9	3
51–60	6,1	0,76	8	2
свыше 60 лет	6,4	0,77	7	1

Для вычисления τ ранжируем группы по K_T в порядке натурального ряда:

Возрастная группа	ранг по X (K_T)	ранг по Y (J_p)	P_i	Q_i
22-24	1	9	0	8
25-30	2	6	2	5
20-21	3	8	0	6
18-19	4	7	0	5
До 18	5	5	0	4
31–40	6	4	0	3
Свыше 60	7	1	2	0
51–60	8	2	1	0
41–50	9	3	0	0

P=5 Q=31

Следовательно, $\tau = \dfrac{5-31}{\frac{1}{2}\cdot 9\cdot 8} = -0,72.$

[109]

Заметим, что для нахождения τ достаточно было найти лишь P и применить формулу (II,6,3). Здесь возникает естественный вопрос: как оценить это значение τ. Ясно, что связь отрицательная (обратная), но насколько значима она?

Проверка существенности. Зададимся вопросом: какова существенность полученного на опыте значения коэффициента корреляции рангов τ или, другими словами, придан ном τ с какой степенью надежности можно утверждать, что связь между двумя признаками действительно существует?

Предположим, что связи нет. Это означает, что, например, при фиксированной последовательности Y-рангов объекта появление любой X-последовательности равновозможно. Объекты всегда можно переставить так, чтобы Y-последовательность оказалась упорядоченной в виде натурального ряда: 1, 2, ..., *N*. Всего различных X-последовательностей (*N*!). Каждая, таким образом, имеет вероятность появления $\frac{1}{N!}$. Каждой X-последовательности соответствует некоторое S = P – Q (и τ, заключенное между –1 и +1). Среди этих τ не все будут различными (см. ниже). Совокупность τ вместе с соответствующими частотами их появления образует некоторое распределение. В дальнейшем, однако, нам будет удобно рассматривать распределение частот S (разумеется, идентичное распределению τ, т.к. τ отличается от S лишь постоянным множителем C_N^2, не меняющим распределение).

Если, например, N = 4, то при заданной Y-последовательности 1,2,3,4 возможны 4! = 1 • 2 • 3 • 4 = 24 X-последовательности (полезно расписать их).

Покажем, что не все они различны (в смысле S) и найдем распределение частот:

$$S = P - Q = 2P - \frac{1}{2}(N-1)N$$

Среди 24-х перестановок найдется лишь одна (4, 3, 2; 1) с *P* = 0 (и S = – 6 соответственно), три (4, 3, 1, 2; 4, 2, 3, 1; 3,4, 2, 1)с P = 1 (S= – 4), пять (4, 2, 1,3; 4, 1,3,2; 3, 4, 1, 2; 3,2, 4, 1; 2, 4, 3, 1) с P = 2 (S = – 2), шесть с P = 3 (S = 0), пять с P = 4 (S = 2), три с P = 5 (S = 4), одна с P = 6 (S = 6).

Таким образом, мы имеем 7 различных S (и τ) с симметричным распределением частот:

[110]

P	0	1	2	3	4	5	6
S	–6	–4	–2	0	2	4	6
n_S	1	3	5	6	5	3	1

$(\sum_s n_S = 24)$

Аналогично можно получить распределения и для других N. Например, для N = 8 число различных S равно 15: O ± 2 ± 4 ± ... ±28. Приведем частоты для S ≥ 0 (для S < O частоты те же, что для S > 0 при одинаковых модулях):

S	n_S	S	n_S	S	n_S
0	3826	10	1940	20	174
2	3736	12	1415	22	76
4	3450	14	961	24	27
6	3017	16	602	26	7
8	2493	18	343	28	1

Максимальная частота соответствует S = 0, с ростом $|S|$ частоты монотонно уменьшаются, достигая 1 при $|S|_{max} = C_N^2$; (|τ| = 1). Если N нечетно, то, оказывается, имеются 2 максимума, приходящиеся на S = ± 1 с увеличением |S| частоты также уменьшаются.

Пусть *N = 3,* имеем 6 перестановок:

1) 3 2 1 *P* = 0 S = – 3 n_S= 1
2) 3 1 2 *P* = 1 S = – 1 n_S= 2
3) 2 3 1 *P* = 1
4) 2 1 3 *P* = 2 S = + 1 n_S= 2
5) 1 3 2 *P* = 2
6) 1 2 3 *P* = 3 S = 3 n_S= 1

Упражнение 51. Для случая *N* = 5 убедиться в справедливости того, что имеются 2 максимума (S = ±1), а с увеличением |S| частота уменьшается, достигая 1 при $|S| = C_N^2$

Уже из рассмотрения случаев N = 4, 5, 8 ясно, что основная часть значений S (и τ) концентрируется вблизи нуля. Если некоторое значение S достаточно далеко от среднего (нулевого), то и вероятность его появления очень мала.

Пример 21. Пусть при *N = 8* значение S = 18 имеет частоту n_S = 343. Вычислим вероятность того, что значение S = 18 появится случайно, т.е. с какой вероятностью мы отвергаем гипотезу независимости (и утверждаем наличие связи).

Событию «S не меньше 18» благоприятствуют 343 + 174 + 76 + 27 + 7 + 1 = 628 равновозможных элементарных событий, следовательно, вероятность равна 628/8! ≈ 0.016, она невелика.

[111]

Обычно используют следующий критерий существенности: если наблюдаемое значение S таково, что вероятность появления этого или большего по абсолютной величине значения достаточно мала (в социальных исследованиях, как уже отмечалось, малой считают вероятность 0,05, а очень малой 0,01), то гипотеза независимости отвергается. Это значит, что S – в «хвостах» распределения. Когда говорят, что .«наблюденное S лежит вне 5-процентного предела существенности», то имеют в виду, что вероятность появления равного или большего по абсолютной величине значения меньше, чем 0,05. (К этому вопросу мы вернемся в главе V).

В нашем примере *(N* = 8, S = 18, $\tau = 0{,}64$) вероятность того, что $|S| \geq 18$, равна $2 \cdot 0{,}016$, следовательно, с надежностью, не меньшей 0,968, можно считать, что между *X* и *Y* есть положительная связь.

Допустим, что для N = 10 $\tau = -0{,}16$. Является ли это значение τ существенным? В данном случае S = – 7. Вероятность того, что $S \leq -7$, как видно из таблицы[19] Г Приложения 3, равна $0{,}30 > 0{,}05$[20]. Мы не можем отвергнуть гипотезу независимости и считать отрицательную связь установленной.

Для N = 10 и $\tau = 0{,}51$ (S = + 23) вероятность того, что $S > 23$, равна (см. таблицу Г) 0,023, а вероятность того, что $|S| > 23$, равна 0,046. Обе вероятности меньше 0,05. Гипотезу о независимости можно отвергнуть с большой надежностью (не меньшей, чем 0,95).

Упражнение 52. Для *N* = 9 и $\tau = -0{,}72$ рассмотреть вопрос о существенности τ. *Ответ:* с надежностью, большей 0,99 гипотеза независимости отвергается.

Упоминавшаяся таблица существенности составлена лишь для $N \leq 10$. Оказывается, что для $N > 10$ нет нужды создавать специальные таблицы. Можно показать, что с ростом *N* очертания полигона частот приближаются к хорошо изученной в статистике кривой нормального распределения (см. (1,3,4)) для

$\sigma^2 = (1/18)N(N—1\ )(2N+5)$

Поэтому можно использовать так называемую таблицу площадей под нормальной кривой[21] (см. § 8 главы V, а также таблицу А Приложения 3).

[112]

[19] Эта таблица построена на основе расчетов, аналогичных тем, которые выполнены в предыдущем примере (для разных *N* и S).

[20] Легко понять, что вероятность $|S| \geq 7$ равна $2 \cdot 0{,}300 = 0{,}600$.

[21] При отсутствии объединенных рангов существенность τ определяется непосредственно по значению τ по таблице Д Приложения 3.

Познакомимся с еще одной формой записи коэффициента Кендэла. Пусть каждый из *N* изучаемых объектов может быть охарактеризован по степени интенсивности как признака X, так и признака Y, т.е. мы знаем у каждого объекта ранг по X и ранг по Y.

Введем величину

$$a_{rs} = \begin{cases} 1, если R_r^{(x)} \succ R_s^{(x)} \\ -1, если R_r^{(x)} \prec R_s^{(x)} \end{cases}$$

где $R_r^{(x)}$ – ранг по X r-ого объекта, а $R_s^{(x)}$ – s-ого. Аналогично вводится величина b_{rs} для признака Y. Станем сопоставлять пары объектов и вычислять произведение $a_{rs} \cdot b_{rs}$. Если большему рангу по X соответствует больший по Y (или меньшему – меньший), то это произведение будет равно 1, так как при этом $a_{rs} = b_{rs} = 1$ (либо $a_{rs} = b_{rs} = -1$). В противном случае (большему рангу по X соответствует меньший по Y или наоборот) произведение $a_{rs}\, b_{rs} = -1$.

Завершив всевозможные сравнения пар элементов, составим сумму соответствующих произведений $S = \sum_r \sum_s a_{rs} \times b_{rs}$. Чтобы одну и ту же пару объектов не сопоставлять дважды, мы будем осуществлять суммирование по *r,* скажем, от 1 до *N*, но тогда по s от $r + 1$ до *N*, т.е. по $s > r$.

Нетрудно видеть, что $S > 0$, если связь прямая и $S < O$, если обратная. S близко к 0, если связи нет. Сконструируем величину

$$\tau = \frac{\sum_{r=1}^{N} \sum_{s=r+1}^{N} a_{rs} b_{rs}}{\sqrt{\sum_{r=1}^{N} \sum_{s=r+1}^{N} a_{rs}^2 \cdot \sum_{r=1}^{N} \sum_{s=r+1}^{N} b_{rs}^2}} \qquad \text{(II,6,4)}$$

Найдем максимальное значение числителя. Оно достигается тогда, когда все $a_{rs} \cdot b_{rs} = 1$. При этом $\tau_{max} = +1$ ($a_{rs}^2 = b_{rs}^2 = 1$).

Аналогично $\tau_{min} = -1$.

Вычислим $\sum\sum a_{rs}^2$. Сопоставление каждого из *N* элементов с другими породит $N - 1$ единицу ($a_{rs}^2 = 1$). Всего таких единиц будет $\frac{1}{2} N(N-1)$. Множитель $\frac{1}{2}$ появляется из-за того, что при такой схеме подсчета каждая пара

[113]

элементов сравнивается дважды. Таким образом $\sum\sum a_{rs}^2 = \sum\sum b_{rs}^2 = \frac{N(N-1)}{2}$. Следовательно,

$$\tau = \frac{\sum_{r=1}^{N}\sum_{s=r+1}^{N} a_{rs} \cdot b_{rs}}{\frac{1}{2}N(N-1)}.$$

Числитель можно несколько упростить.

Расположим объекты по рангу X, тогда все $a_{rs} = 1$. При этом

$$\sum_{r=1}^{N}\sum_{s=r+1}^{N} a_{rs} \cdot b_{rs} = \sum_{r=1}^{N}\sum_{s=r+1}^{N} b_{rs} = P - Q,$$

где P, очевидно, получим, суммируя числа, показывающие, сколько рангов образовавшегося рангового ряда Y превышают ранги, занимаемые первым, вторым и т.д. *N*-ным, а Q – аналогичная сумма, показывающая, сколько рангов ряда Y ниже рангов, записанных первым, вторым и т.д. *N*-ным. Таким образом, приходим к уже известному коэффициенту: см. (II,6,1).

Итак, мы познакомились с новой формой записи коэффициента Кендэла (II,6,4).

Далее, допустим, что *t* рангов по X с *l*+ 1 по *l* + *t* объединены, т.е. ранговый ряд имеет вид:

$$1,2,...,l,l+\frac{1+t}{2},l+\frac{1+t}{2},...l+\frac{1+t}{2},l+t+1,...,N$$

Сопоставление всех не объединенных рангов с другими, объединенными и не объединенными, дадут те же результаты, что и ранее: в нашем примере ранг объединенных все равно выше рангов 1, 2, ..., *l* и ниже рангов *l*+ *t*+ 1, ..., *N*. Но сопоставление объединенных рангов между собой не будет порождать ни +1, ни –1, так как эти ранги равны. Доопределим теперь a_{rs} и b_{rs}, так, чтобы $a_{rs} = b_{rs} = 0$ при совпадении рангов (это естественно). Всего сопоставлений объединенных рангов $\frac{t(t-1)}{2}$. Сумма $\sum_r\sum_s a_{rs}^2$ уменьшится на $\frac{t(t-1)}{2}$. Если объединений несколько, скажем, p, а t_v – число объединенных рангов в *v*-ом объединении по X,

[114]

то сумма уменьшится на величину

$$U_x = \sum_{v=1}^{p} \frac{t_v(t_v - 1)}{2}$$

Пусть q – число объединенных рангов y, а u_w – число объединенных рангов в *w*-ом объединении, тогда сумма $\sum\sum b_{rs}^2$ уменьшится на

$$U_y = \sum_{w=1}^{q} \frac{u_w(u_w - 1)}{2}$$

Итак, для случая объединенных рангов окончательно имеем:

$$\tau = \frac{P - Q}{\sqrt{(\frac{N(N-1)}{2} - U_x)(\frac{N(N-1)}{2} - U_y)}}. \text{ (II,6,5,)}$$

В отличие от ρ коэффициент τ без поправки меньше, чем коэффициент τ с поправкой, т.е. использование τ без поправок повышает ошибку II рода и менее опасно, чем использование ρ без поправок (см. гл. V).

Пример 22. Рассмотрим следующую таблицу:

Объекты	A	B	C	D	E	F	G	H	K	L	M	N
X	1,5	1,5	3	4	6	6	6	8	9,5	9,5	11	12
Y	2,5	2,5	7	4,5	1	4,5	6	11,5	11,5	8,5	8,5	10

Что порождает в S элемент A?

При сопоставлении A с S, очевидно, 0 (одинаковые ранги по X), A с C – плюс единицу (+1) × (+1) = 1, аналогично 1 порождает сопоставление A с D, F, G, H, K, L, M, N; при сопоставлении A с E появляется минус единица (ранг по X в прямой, а по Y – в обратной последовательности: 1 × (–1) = –1).

Таким образом, вклад A в S равен +8. Продолжая сопоставления, получим: S = 8 + 8 + 1 +5 + 5 + 5 + 5 – 3 – 2 + 1 + 1 = 34.

В X – последовательности три объединения: $t_1 = 2$; $t_2 = 3$; $t_3 = 2$; $U_x = 5$; во второй – четыре: $u_1 = u_2 = u_3 = u_4 = 2$; $U_y = 4$. Теперь по формуле (II,6,5): τ = 0,55.

Упражнение 53. В упоминавшейся книге «Методика и техника статистической обработки первичной социологической информации» приводится таблица «Вычисление

[115]

коэффициента корреляции рангов Кендэла между ответами рабочих: «интересная работа» и «образование соответствует работе» (с. 17). Воспроизведем часть ее.

Рассчитать τ. В случае необходимости помочь в этом может цитируемая книга. Там, в частности, показывается, что

Таблица 28

Пример вычисления коэффициента ранговой корреляции Кендэла

Номер профессиональной группы	X –, ответившие, что работа интересная, %	ранг по X	У – лица, ответившие, что образование соответствует работе, %	ранг по У
1	100,0	3	100	1
2	100,0	3	87,5	5,5
3	100,0	3	77,0	9
4	100,0	3	75,0	10
5	100,0	3	50,0	11,5
6	83,5	6,5	92,0	3
7	83,5	6,5	83,5	8
8	83,0	8	90,0	4
9	82,5	9	94,5	2
10	71,0	10	87,0	7
11	55,5	11	87,5	5,5
12	50,0	12	50,0	11,5
13	28,5	13	43,0	13
14	0	14	0	14

P = 61, Q = 28, однако при вычислении τ не учтено, что имеются объединения рангов. Даже если Вы используете книгу, рассчитайте τ самостоятельно, с учетом объединений. Для контроля: $U_x = 1$, $U_y = 2$. Ответ: $\tau = +0{,}39$.

Об оценке существенности τ в случае объединенных рангов см. § 8 главы V.

До сих пор использовались формулы, справедливые для любых *N*, однако удобные лишь для малых (не более 20–30); в противном случае вычисления существенно затрудняются.

Сейчас мы рассмотрим большие *N*. В таких случаях признаки шкалируются. Как и ранее, будем считать, что признак X принимает значения x_i где $i = \overline{1,k}$, а признак Y – значения y_j, где $j = \overline{1,l}$ (обычно *k*, *l* ≈ 5–10). Эмпирический материал сводится в корреляционную таблицу $\{N_{ij}\}$, для которой $\sum_i \sum N_{ij} = N$ (см. § 1, главы II).

[116]

В качестве исходной возьмем формулу

$$\tau = \frac{S}{\sqrt{A \cdot B}},\ A = \sum_r \sum_s a_{rs}^2,\ B = \sum_r \sum_s b_{rs}^2$$

$$S = \sum_r \sum_s a_{rs} b_{rs}\ . \qquad \text{(II,6,6)}$$

При больших N выполнить суммирование по r и s от 1 до N чрезвычайно затруднительно, поэтому перейдем к суммированию по i и j от 1 до k и l соответственно.

Рассмотрим A. Нам нужно сравнить ранги по X каждой пары объектов, а результаты просуммировать[22]. Очевидно, можно не сравнивать между собой элементы строки, так как у них одинаковые ранги по X. Следовательно, все элементы, у которых $X = x_1$ (всего их $N(x_1)$), можно не сравнивать друг с другом, но следует сравнить с элементами, у которых $X = x_2$. Такое сравнение породит $N(x_1) \cdot N(x_2)$ единиц, а сравнение элементов с $X = x_1$ с элементами, у которых $X = x_3$, дает $N(x_1)\,N(x_3)$ единиц и т.д. Поэтому

$$A = N(x_1)\,[N(x_2) + N(x_3) + \ldots + N(x_k)] + N(x_2)\,[N(x_3) + N(x_4) + \ldots + N(x_k)] + \ldots + N(x_{k-1})N(x_k) =$$

$$A = N(x_1)\left[N(x_2) + N(x_3) + \ldots + N(x_k)\right] + N(x_2)\left[N(x_3) + N(x_4) + \ldots + N(x_k)\right] + \ldots +$$

$$+ N(x_{k-1})N(x_k) = \sum_{i=1}^{k-1} N(x_i) \sum_{p=1}^{k-1} N(x_{i+p}) \qquad \text{(II,6,7)}$$

Упражнение 54. Показать, что

$$B = \sum_{j=1}^{l-1} N(y_j) \sum_{q=1}^{l-j} N(y_{j+q}) \qquad \text{(II,6,8)}$$

Перейдем к рассмотрению S. Теперь для каждой пары элементов нужно сравнивать и ранги по X (a_{rs}), и ранги по Y (b_{rs}).

Рассмотрим элементы клетки (i, j). Ясно, что их не нужно сравнивать ни с элементами i-ой строки (об этом мы уже говорили), ни с элементами j-го столбца (у элементов столбца одинаковые ранги по Y, следовательно, за счет b_{rs} соответствующее слагаемое обратится в нуль). Станем сравнивать некоторый элемент из клетки (i, j) с элементом клетки (i', j'), если $i' > i$, $j' > j$. Такое сравнение для каждой пары объектов породит +1 в силу упорядоченности пунктов шкалы ($a_{rs} = 1$, $b_{rs} = 1$). Если $i' > i$, а $j' > j$, то каждая пара породит –1 ($a_{rs} = 1$, $b_{rs} = -1$). Суммируя по i',j', мы

[117]

[22] В дальнейшем изложении предполагается, что значения X и Y выписаны в таблице в порядке возрастания (сверху вниз и слева направо).

переберем всевозможные сравнения выделенного элемента из клетки (i, j) со всеми элементами, лежащими ниже и справа ($j'> j$, $i'>i$) которые дадут, таким образом, $\sum_{i'=i+1}^{k}\sum_{j'=j+1}^{l} N_{i'j'}$. Сопоставление элемента из клетки (i,j) с элементами, расположенными ниже и слева от этой клетки, порождает слагаемое $\sum_{i'=i+1}^{k}\sum_{j'=1}^{j-1} N_{i'j'}$. Так как все элементы клетки (i,j) равно-

Таблица 29

Связь удовлетворенности работой с удовлетворенностью специальностью

X	Y			N(x_i)
	удовлетворен	промежуточная позиция	не удовлетворен	
удовлетворен	1472	50	65	1587
промежуточная позиция	136	65	42	243
не удовлетворен	126	42	165	333
N(y_j)	1734	157	272	2163

правны, то умножая результат на N_{ij} и суммируя затем по i и j, мы осуществим вообще все возможные сравнения пар элементов.

Упражнение 55. Почему не нужно рассматривать случай $i'<i$?

Итак,

$$S = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}\left(\sum_{i'=i+1}^{k}\sum_{j'=j+1}^{l} Ni'j' - \sum_{i'=i+1}^{k}\sum_{j'=1}^{j-1} Ni'j'\right) \quad \text{(II,6,9)}$$

Тем самым мы завершили переход к корреляционной таблице во всех множителях τ[23].

Для иллюстрации этой «страшной» формулы приведем пример, который покажет справедливость пословицы «не так страшен черт, как его рисуют».

Пример 23. Изучая связь удовлетворенности работой (Y) с удовлетворенностью специальностью (X) мы, в частности, получили корреляционную таблицу 29 (массив, ОСРЗ).

[118]

[23] Авторы выражают благодарность Г.И. Саганенко за помощь при выводе соотношения (II,6,9).

Теперь A = 1587 (243 + 333) + 243·333 = 995031;

B = 1734 (157 + 272) + 157·272 = 786590;

S = 1472 (65 + 42 + 42 + 165) + 50 (42 + 165 – 136 – 126) – 65 (136 + 65 + 126 + 42) + 136 (42 + 165) + 65 (165 – 126) – 42 (126 + 42) = 459104;

$\tau = +0{,}52$.

Таким образом, между изучаемыми удовлетворенностями есть тесная положительная связь.

Упражнение 56. Для признаков удовлетворенность работой (Y), удовлетворенность общественной работой (X) корреляционная таблица имеет вид:

Таблица 30

Связь удовлетворенности работой (Y) с удовлетворенностью общественной работой (X)

X	Y			$N(x_i)$
	Y_1	Y_2	Y_3	
x_1	1241	82	150	1473
x_2	147	11	38	196
x_3	103	13	13	129
N(yj)	1491	106	201	1798

Вычислить τ. Ответ: $\tau = +0{,}31$.

Связь, таким образом, тоже положительная, но менее тесная. Еще менее тесной, например, оказывается связь между удовлетворенностью работой и удовлетворенностью досугом (для соответствующей корреляционной таблицы $\tau = +0{,}14$), что допускает естественную интерпретацию.

Коэффициент τ, определяемый формулой (II,6,6), может обращаться в ±1 только в том случае, когда таблица диагональна.

В самом деле, согласно неравенству Коши[24] |S| максимален, если наборы a_{rs} и b_{rs} пропорциональны: $b_{rs} = \alpha \cdot a_{rs}$. Это возможно лишь тогда, когда все наблюдения либо на положительной ($\alpha = 1$), либо на отрицательной ($\alpha = -1$) главной диагонали таблицы, т.е. если таблица квадратная (если есть не диагональные элементы, то α не будет знако-

[119]

[24] Для читателя, незнакомого с этим неравенством, мы приводим его вывод в конце параграфа.

постоянной величиной, соотношение $b_{rs} = \alpha\, a_{rs}$ не будет выполняться для всех пар элементов).

Для прямоугольной таблицы |S| достигает максимума, если: 1) все наблюдения лежат в клетках самой длинной диагонали таблицы, т.е. диагонали, содержащей m = *min* (*k, l*) клеток, так как в случае появления недиагональных элементов в S, кроме нулей типа 0·0, добавляются нули типа $a_{rs} \cdot 0$ и $0 \cdot b_{rs}$, причем за счет уменьшения числа слагаемых, равных 1;

2) все наблюдения равномерно распределены между диагональными клетками, т.е. $N_{ii} = N/m$ (так как обычно N >> m, то можно считать, что оно кратно m без существенной потери точности).

Проиллюстрируем первое утверждение, например, для следующей таблицы:

X	Y		N(x_i)
	y_1	y_2	
x_1	N_{11}	1	N_{11}+1
x_2	0	$N_{22} - 1$	$N_{22} - 1$
x_3	0	0	0
N(y_j)	N_{11}	N_{22}	N

$$S = N_{11}(N_{22} - 1) \prec N_{11}N_{22}$$

Проиллюстрируем второе утверждение. Рассмотрим, например, диагональную таблицу 3 × 3:

X	Y			N(x_i)
	y_1	y_2	y_3	
x_1	N_{11}	0	0	N_{11}
x_2	0	N_{22}	0	N_{22}
x_3	0	0	N_{33}	N_{33}
N(y_j)	N_{11}	N_{22}	N_{33}	N

Для нее

$$S = N_{11}N_{22} + N_{11}N_{33} + N_{22}N_{33} \le N_{11}^2 + N_{22}^2 + N_{33}^2,$$

$$S_{\max} = N^2/3 \textit{при} N_{11} = N_{22} = N_{33} = N/3,$$

т.е.

[120]

если все наблюдения распределены равномерно. Здесь мы использовали известное неравенство

$ab + bc + ac \le a^2 + b^2 + c^2$,

которое легко получить, складывая почленно три очевидных неравенства

$(a-b)^2 \ge 0, (a-c)^2 \ge 0, (b-c)^2 \ge 0$.

В общем случае в каждой клетке самой длинной диагонали должно быть N/m элементов. Сопоставляя элементы первой клетки с остальными, мы получим $\frac{N}{m} \cdot \frac{N}{m}(m-1)$ единиц, а элементы второй с прочими $\frac{N}{m} \cdot \frac{N}{m}(m-2)$, так как их уже не нужно сравнивать с элементами первой и т.д.

В итоге

$$S_{\max} = \frac{N^2}{m^2}\left[(m-1) + (m-2) + \cdots + 2 + 1\right] = \frac{N^2(m-1)}{2m}$$

Но при этом значении S коэффициент τ, вообще говоря, не достигает значений ± 1.

Введем

$$\tau_c = \frac{S}{S_{\max}} = \frac{2mS}{N^2(m-1)}. \qquad \text{(II,6,10)}$$

Очевидно, он принимает значения, которые могут достичь ± 1 (если не считать незначительного эффекта, возникающего в случае, когда N не кратно *m)* даже для прямоугольных таблиц.

Коэффициент, определяемый (II,6,6), обозначают иногда τ_b, а (II,6,1) – τ_a, если нет объединений рангов $\tau_a = \tau_b$.

Обратим внимание на то, что три коэффициента r, ρ, τ можно рассмотреть с единой точки зрения. Действительно, пусть, как обычно, имеется совокупность из *N* индивидов, каждый из которых может быть охарактеризован с помощью значений двух признаков *X* и *Y*.

Выберем пару индивидов, например, *i* и *j* и станем приписывать ей некоторую *x* – оценку a_{ij} (конкретизация оценок будет дана ниже), обладающую свойством антисимметричности: $a_{ij} = -a_{ij}$ Аналогично введем *y* – оценку b_{ij}.

[121]

Рассмотрим величину

$$\Gamma = \frac{\sum_i \sum_j a_{ij} b_{ij}}{\sqrt{\sum_i \sum_j a_{ij}^2 \cdot \sum_i \sum_j b_{ij}^2}}$$

Мы уже видели (II,6,6), что для величины

$$a_{ij} \begin{cases} 1, если R_i^{(x)} \succ R_j^{(x)} \\ -1, если R_i^{(x)} \prec R_j^{(x)} \end{cases}$$

(где $R_i^{(x)}$ – ранг по X *i*-го элемента) и аналогичной величины b_{ij}: $\Gamma = \tau$.

Пусть

$a_{ij} = x_j - x_i$, а, $b_{ij} = y_j - y_i$

тогда

$$\sum_i \sum_j (x_j - x_i)(y_j - y_i) = 2N \sum_i x_i y_i - 2 \sum_i \sum_j x_i y_j$$

$$\sum_i \sum_j (x_j - x_i)^2 = 2N \sum_i x_i^2 - 2(\sum_i x_i)^2$$

Теперь

$$\Gamma = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

Если положить $a_{ij} = R_j^{(x)} - R_i^{(x)}$, а $b_{ij} = R_j^{(y)} - R_i^{(y)}$, то можно аналогично предыдущему показать, что Γ обращается при этом в ρ. Это рассмотрение составит для читателя самостоятельное *упражнение 57*.

Мы же сошлемся на § 5 главы II, где было показано, что ρ является r, примененным к рангам, а так как для r рассмотрение проведено, то с точки зрения строгости изложения, выкладки данного упражнения в тексте книги не являются необходимыми. В заключение выведем неравенство Коши.

Очевидное неравенство $(A_{ij} - B_{ij})^2 \geq 0$ можно переписать в виде $\frac{1}{2} A_{ij}^2 + \frac{1}{2} B_{ij}^2 \geq A_{ij} B_{ij}$

Полагая

$$A_{ij} = \frac{a_{ij}}{\sqrt{\sum \sum a_{ij}^2}} \text{ и } B_{ij} = \frac{b_{ij}}{\sqrt{\sum \sum b_{ij}^2}}$$

[122]

и суммируя всевозможные неравенства, получим:

$$\frac{1}{2}\frac{\sum_i\sum_j a_{ij}^2}{\sum_i\sum_j a_{ij}^2}+\frac{1}{2}\frac{\sum_i\sum_j b_{ij}^2}{\sum_i\sum_j b_{ij}^2}\geq\frac{\sum_i\sum_j a_{ij}b_{ij}}{\sqrt{\sum_i\sum_j a_{ij}^2\sum_i\sum_j b_{ij}^2}}$$

Так как левая часть равна 1, то неравенство Коши доказано. Нетрудно видеть, что оно превращается в равенство, если все $a_{ij} = \alpha b_{ij}$ (убедиться подстановкой!), что и было нами ранее использовано.

Наконец, рассмотрим случай, когда оба признака измерены на уровне наличия – отсутствия.

Пусть индекс 1 соответствует наличию, а 2 отсутствию признака, тогда корреляционная таблица для признаков X и Y принимает вид:

X	Y		$N(x_i)$
	y_1	y_2	
x_1	N_{11}	N_{12}	$N(x_1)$
x_2	N_{21}	N_{22}	$N(x_2)$
$N(y_j)$	$N(y_1)$	$N(y_2)$	N

Каждый элемент первой клетки положительной диагонали при сопоставлении с элементом второй породит +1, всего таких +1 в S войдет $N_{11}\cdot N_{22}$.

Сравнение элементов отрицательной диагонали породит $N_{12}\cdot N_{21}$, отрицательных единиц.

Следовательно,

$S = N_{11}N_{22} - N_{12}N_{21}$;

$U_x = \frac{1}{2}N(x_1)\left[N(x_1)-1\right]+\frac{1}{2}N(x_2)\left[N(x_2)-1\right]$; а

$\frac{1}{2}N(N-1)-U_x = N(x_1)N(x_2)$

Аналогично:

$\frac{1}{2}N(N-1)-U_y = N(y_1)N(y_2)$

[123]

теперь коэффициент Кендэла, определяемый (II,6,5):

$$\tau = \frac{N_{11}N_{22} - N_{12}N_{21}}{\sqrt{N(x_1)N(x_2)N(y_1)N(y_2)}}$$

таким образом, совпадает с коэффициентом Ф (II,3,2).

Этот результат проясняет смысл формально введенного ранее коэффициента контингенции.

7. Энтропийные меры в социологическом анализе

Пусть некоторое событие может иметь k различных исходов $A_i (i = \overline{1,k})$ вероятность которых обозначим через $P(A_i)$. Ясно, что $\sum_{i=1}^{k} P(A_i) = 1$. Например, при подбрасывании симметричной монеты $k = 2$, A_1 — выпадение герба, A_2 — решки, $P(A_1) = P(A_2) = \frac{1}{2}$

Допустим, что мы хотим предсказать исход испытания. Если $k = 1$, то исход предопределен. Если $k = 2$, то появляется неопределенность, которая максимальна при $P(A_1) = P(A_2)$. Если $P(A_1) > P(A_2)$, то чем больше $P(A_1)$, тем меньше неопределенность предсказания. В пределе, когда $P(A_1) = 1$ ($P(A_2) = 0$), неопределенность исчезает: во всех испытаниях осуществляется исход A_1.

Чем больше k, тем менее определенны предсказания, тем больше неопределенность. По К. Шеннону, мерой неопределенности является величина $E = -\sum_{i=1}^{k} P(A_i) \log P(A_i)$, называемая *энтропией*. Если неопределенности нет и, скажем, реализуется l-ое состояние, т.е. $P(A_l) = 1$, а все остальные $P(A_i) = 0$, то E очевидно, обращается в нуль. Неопределенность максимальна, если все исходы равновозможны, т.е. $P(A_i) = 1/k$. При этом $E_{\max} = \log k$. Чем больше k, тем больше $E_{\max}$. Итак, $0 \le E \le \log k$.

Пусть N индивидов некоторой совокупности обладают некоторым признаком X, и событие A_i состоит в том, что значение признака равно x_i. Обозначим через N_i число индивидов, у которых $X = x_i$. Если N достаточно велико, то $P_i = N_i/N$, а E – мера «распыленности» распределения. Для сопоставления различных распределений целесообразно перейти к нормированному коэффициенту $\varepsilon = E/E_{max}$. Величина ε, принимающая значения между 0 и 1, является *аналогом дисперсии*.

[124]

Перейдем к двухмерным распределениям для признаков X и Y в случае, когда эмпирический материал сведен в корреляционную таблицу $\{N_{ij}\}$.

Теперь

$$E = -\sum_{i=1}^{k}\sum_{j=1}^{l} P_{ij} \log P_{ij} ,$$

где $P_{ij} = N_{ij}/N$ и суммирование ведется по всем клеткам корреляционной таблицы. Здесь и далее мы не указываем основание логарифма, так как обсуждаемые относительные показатели ε и λ, от него не зависят.

Упражнение 58. Показать, что $E_{max} = \log kl$

Упражнение 59. Показать, что теперь

$$\varepsilon = \frac{N \log N - \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij} \log N_{ij}}{N \log kl}$$

Это выражение используется для расчета *энтропийной меры дисперсии*

Рассмотрим теперь так называемую *энтропийную меру связи*. Неопределенность Y-распределения

$E_y = -\sum_{j=1}^{l} \frac{N(y_j)}{N} \log \frac{N(y_j)}{N}$, если ничего не известно об X-распределении.

Неопределенность Y-распределения у индивидов с $X = x_i$, так называемая условная неопределенность

$$E_{y/x_i} = -\sum_{j=1}^{l} \frac{N_{ij}}{N(x_i)} \log \frac{N_{ij}}{N(x_i)} \quad (i = \overline{1,k})$$

В итоговую условную неопределенность каждая строчка таблицы дает вклад с удельным весом $N(x_i)/N$, т.е. полная условная неопределенность Y-распределения:

$$E_{y/x} = \sum_{i=1}^{k} \frac{N(x_i)}{N} E_{y/x_i}$$

Мерой связи между признаками X и Y может служить величина относительной неопределенности

$$\lambda_{y/x} = \frac{E_y - E_{y/x}}{E_y} .$$

[125]

Упражнение 60. Рассмотреть для простейших таблиц 2x2 случай отсутствия связи и показать, что $\lambda = 0$. *Указание*: использовать, что $N_{ij} = N(x_i)N(y_j)/N$.

Упражнение 61. Рассмотреть случаи функциональной связи и показать, что $\lambda = 1$. *Указание*: учесть, что таблица принимает диагональный вид. Итак, $0 \le \lambda \le 1$. Чем больше λ, тем больше связь между признаками.

Упражнение 62. Вычислить ε и $\lambda_{y/x}$ для следующей таблицы:

X	Y						$N(x_i)$
	y_1	y_2	y_3	y_4	y_5	y_6	
x_1		14	28	48	66	45	202
x_2	1	35	53	40	36	8	173
x_3	3	39	15	13	4	2	76
$N(y_j)$	5	88	96	101	106	55	451

Ответ: $\varepsilon = 0{,}872$, $Ey/x_1 = 0{,}664$, $Ey/x_2 = 0{,}660$, $Ey/x_3 = 0{,}582$, $\lambda_{y/x} = 0{,}086$.

Упражнение 63. Обратимся к рассмотрению связи между удовлетворенностью работой и удовлетворенностью заработной платой. Для таблицы 18 (работники в возрасте до 30 лет) найти λ.

Ответ: $Ey/x_1 = 0{,}289$, $Ey/x_2 = 0{,}396$, $Ey/x_3 = 0{,}383$, $Ey/x = 0{,}348$, $\lambda = 0{,}030$.

Упражнение 64. Для таблицы 19 (работники старше 30 лет) найти λ. Ответ: $\lambda = 0{,}014$.

Таким образом, связь между рассматриваемыми показателями более тесная для молодых работников. В дальнейшем мы вернемся к этому вопросу еще раз, используя другие методы статистического изучения связей (§ 8 главы II).

Пример 24. Представляет несомненный интерес задача о связи интегральной удовлетворенности с частными удовлетворенностями (отдельными элементами рабочей ситуации).

В качестве элементов обычно выделяют: 1) содержание труда (совокупность трудовых функций, выполняемых в процессе создания потребительных стоимостей в процессе труда), 2) условия (факторы, под воздействием которых осуществляется трудовая деятельность: сменность, физическая нагрузка, состояние окружающей среды и т.д.);

[126]

3) организация (совокупность мероприятий, обеспечивающих рациональное использование рабочей силы); 4) оплата; 5) межличностные отношения и т.д.

Осознавая, что человек не может точно определить вклад, который вносит в общее состояние удовлетворенности удовлетворение отдельных потребностей, мы отказались от метода ранжирования различных факторов. Для изучения обсуждаемой связи использовались различные статистические показатели, которые вычислялись для распределений совокупности в случае, когда одним из признаков является интегральная удовлетворенность и другим – последовательно-частные.

Для T и λ элементы расположились так: содержание труда, организация, оплата, отношения с администрацией и т.д. (см. также § 8 гл. II). Заметим, что при интерпретации следует учитывать, что рассматриваемые элементы не являются независимыми: содержание труда, например, нельзя считать «очищенным» от влияния зарплаты, ибо в среднем более содержательная работа выше оплачивается и т.д. Следует также учитывать, что речь идет об *оценках* элементов, а связь между элементом и оценкой носит сложный, опосредствованный характер. Например, нет прямой зависимости между удовлетворенностью зарплатой и ее величиной (в наших исследованиях было установлено наличие U-образной зависимости[25]). Зависимости опосредствуются потребностями, притязаниями. Так, удовлетворенность зарплатой зависит не столько от ее «абсолютной» величины, сколько от достижения «нормы», в качестве которой, как удалось установить, выступает средняя прогрессивная референтной группы (для работников промышленных предприятий ею оказалась их социально-профессиональная группа). Во всяком случае нами установлена тесная корреляция между удовлетворенностью зарплатой рабочих данной группы и числом работников, получающих зарплату не ниже среднепрогрессивной[26].

Пример 25. Коэффициент λ, определенный выше, описывает влияние X на Y. Мы обозначим его $\lambda_{y/x}$. Аналогично

[127]

[25] Аналогичный характер имеет зависимость между удовлетворенностью образованием и фактическим образованием (обследовались работники промышленных предприятий г. Одессы).

[26] *Максименко В. С., Попова И. М.* Заработная плата как фактор стимулирования трудовой деятельности.— В кн.: Проблемы экономики моря и мирового океана. Одесса, 1973, № 2

можно ввести коэффициент $\lambda_{x/y} = \frac{E_x - E_{x/y}}{E_x}$, который описывает влияние Y на X. λ несимметричен: вообще говоря $\lambda_{y/x} \neq \lambda_{x/y}$ Если из содержательного анализа ясно, что X может влиять на Y и Y на X, то целесообразно вычислить оба коэффициента. Например, удовлетворенность работой (Y), влияет на удовлетворенность специальностью (X) и наоборот. Поэтому мы вычисляем оба коэффициента, используя их для сравнения указанных влияний. Так, в конкретном исследовании рабочих Ильичевского судоремонтного завода (1974г.) нами было получено такое двумерное распределение обсуждаемых признаков:

Таблица 31

Связь между удовлетворенностью работой и удовлетворенностью специальностью

X	Y			$N(x_i)$
	y_1	y_2	y_3	
x_1	1105	30	110	1245
x_2	313	55	62	430
x_3	35	4	36	75
$N(y_j)$	1453	89	208	1750

Для этой корреляционной таблицы, оказывается, $\lambda_{y/x}$ = 0,073, а $\lambda_{x/y}$ = 0,057. Таким образом, можно предположить, что удовлетворенность специальностью в большей мере влияет на удовлетворенность работой (предприятием), чем наоборот. Подчеркнем, что это утверждение относится к локальным условиям определенного, весьма специфического предприятия. Для изучения поставленного вопроса в целом необходимо провести дальнейшие исследования. В нашу задачу здесь входило ознакомление с идеей метода и техникой вычисления.

Упражнение 65. Вычислить $\lambda_{y/x}$ и $\lambda_{x/y}$ для таблицы из упражнения 62 самостоятельно.

Пример 26. *Энтропийный анализ социальных структур*.

В шестидесятые годы О.И. Шкаратан с группой сотрудников изучал социальную структуру современного промышленного предприятия. Результаты теоретического анализа, базирующегося на значительном эмпирическом материале, изложены в книге «Проблемы социальной структуры рабо-

чего класса СССР» (М., 1970). Совместно с И.Н. Тагановым О.И. Шкаратан предпринимал попытки использования количественного метода для изучения указанной структуры. Одна из них, связанная с применением энтропийного анализа, была изложена в журнале «Вопросы философии» (1969, №5) и привлекла внимание социологов, интересующихся использованием количественных методов в социальных исследованиях. Рассмотрим ее суть применительно к фактически реализованной исследователями программе, но с использованием обозначений предыдущих параграфов.

Основная задача, которая решалась авторами с помощью энтропийного анализа, состояла в выделении свойств (признаков), определяющих неоднородность изучаемой социальной структуры. Задача рассматривалась в трехмерном пространстве, т.е. из гипотетического набора значимых признаков (он был составлен на основе предварительного анализа, сюда вошли такие характеристики, как образование, квалификация, пол, профессия и т.д. – всего 27 признаков) авторы выделяли каждый раз тройку признаков, набор которых давал различные пространства. Всего таких пространств можно выделить $C_{27}^{3} = 2925$.

Логика исследования такова. Каждый индивид данной совокупности является носителем различных признаков. Пусть он обладает *i*-м значением признака X, *j*-м – У, *r*-м – Z (в соответствии с ограничением, принятым авторами, мы рассматриваем пространство, определяемое признаками X, Y, Z). Информацию об одном индивиде можно рассматривать как вектор в данном пространстве. Совокупности из N рассматриваемых индивидов соответствует совокупность N векторов. Из всех возможных пространств (наборов признаков) нужно выделить такое, в котором векторы лежат наиболее плотными группами (набор признаков наиболее резко дифференцирует совокупность индивидов). Для отыскания таких пространств и был применен энтропийный анализ.

Неопределенность заполнения пространства векторами определяется величиной

$$E = -\sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{r=1}^{m} P_{ijr} \log P_{ijr} ,$$

где $P_{ijr} = \frac{N_{ijr}}{N}$ (здесь N_{ijr} – число индивидов, у которых X = x_i, Y = y_i, Z = z_i); $i = \overline{1,k}$; $j = \overline{1,l}$; $r = \overline{1,m}$

[129]

Если векторы равномерно заполняют пространство, то

$$N_{ijr} = \frac{N}{klm},\ P_{ijr} = \frac{1}{klm},\ E_{\max} = \log klm$$

Рассмотрим величину $\alpha = \frac{E_{\max} - E}{E_{\max}}$. Так как $0 \le E \le E_{\max}$, то $0 \le \alpha \le 1$, причем $\alpha = 0$ соответствует $E = E_{max}$, т.е. отсутствию неоднородности в распределении векторов (отсутствию дифференциации общности), а $\alpha = 1$ соответствует $E = 0$, т.е. максимальной неоднородности (максимальной дифференциации).

Очевидно, разным пространствам соответствуют различные α и формально задача сводится к отысканию пространства с максимальным α.

Упражнение 66. Показать, что

$$\alpha = \frac{N \log \frac{klm}{N} + \sum_i \sum_j \sum_r N_{ijr} \log N_{ijr}}{N \log klm}$$

На эмпирическом материале ленинградских социологов величина α оказалась максимальной для набора признаков «профессия – квалификация – образование». Именно в этом пространстве векторы лежат наиболее плотными группами, данный набор признаков наиболее резко дифференцирует изучаемую социальную общность.

8. Некоторые другие коэффициенты

В данном параграфе мы рассмотрим несколько статистических коэффициентов, которые не получили в социальных исследованиях такого широкого распространения, как, скажем, r, ρ, T и даже η). Однако в социологической литературе уже встречаются упоминания об их использовании отдельными авторами.

Мы считаем целесообразным рассмотреть определения, проанализировать их и привести примеры вычисления некоторых таких коэффициентов[27]. С одной стороны, это покажет читателю, что диапазон используемых методов значительно шире, чем может представиться по основной массе публикаций, с другой, позволит более свободно ориентироваться в научных статьях.

[130]

[27] Обзор ряда других коэффициентов можно найти в кн.: Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. М., 1977, гл. III, IV.

g – *коэффициент Гудмана (для номинальных шкал)*

Коэффициент Гудмана не является симметричным: $g_{yx} \neq g_{xy}$ Если мы рассматриваем X как независимый (факторный) признак, то его влияние на Y описывается с помощью коэффициента

$$g_{yx} = \frac{\sum_{i=1}^{k} \max N_{ij} - \max N(y_i)}{N - \max N(y_j)}, \text{ (II,8,1)}$$

где *max* $N(y_j)$ – максимальный маргинал зависимого признака, а *max* N_{ij} – максимальная частота в *i*-ой строке корреляционной таблицы.

Если данному X соответствует определенный Y, то в строке лишь одна частота с соответствующим маргиналом, искомая сумма максимумов обращается в N, следовательно, $g_{yx} = 1$

Если признаки независимы, то $N_{ij} = \frac{1}{N} N(x_i) N(y_j)$, как мы видели, и максимальная частота в *i*-ой строке там, где максимален Y-маргинал, т.е.

$$\sum_{i=1}^{k} \max N_{ij} = \frac{\max N(y_j)}{N} \sum_{i=1}^{k} N(x_i) = \max N(y_j)$$

Теперь $g_{yx} = 0$. Итак, $0 \leq g_{yx} \leq 1$. Аналогично определяется g_{xy}, описывающий влияние Y на X.

Коэффициенты Гудмана целесообразно использовать, если из содержательных соображений ясно, что X может влиять на Y (и наоборот) и это влияние, вообще говоря, не симметрично.

В тех случаях, когда X не может влиять на Y (например, X – квалификация, Y – возраст), следует вычислять только g_{xy} (возраст влияет на квалификацию).

Упражнение 67. Рассчитать g_{yx} и g_{xy} для следующей корреляционной таблицы:

X	Y			$N(x_i)$
	y_1	y_2	y_3	
x_1	20	0	0	20
x_2	0	15	30	45
$N(y_j)$	20	15	30	65

Ответ: $g_{yx} = 0{,}57$; $g_{xy} = 1$.

[131]

Интерпретируем результат:

Задание Y однозначно определяет X (см. таблицу). Соответственно $g_{yx} = 1$; но задание X не определяет еще Y (например, если $X = x_2$, то Y может быть и y_2, и y_3), соответственно $g_{yx} < 1$

Упражнение 68. 1. Записать любую диагональную таблицу и убедиться, что $g_{yx} = g_{xy} = 1$.

2. Сконструировать таблицу, для которой $g_{xy} < 1$, а $g_{yx} = 1$, Интерпретировать результаты расчета по аналогии с предыдущим.

Заметим, что выполнение этих несложных упражнений помогает уяснить смысл и различие коэффициентов g_{yx} и g_{xy}. Отметим также предлагаемый Б. Миркиным подход к обработке социологической информации[28], который может быть использован даже для случая номинальных шкал. В качестве меры близости признаков рассматривается мера близости разбиений общности, осуществляемых этими признаками.

Коэффициент близости разбиений

Обобщим формулу для меры близости между двумя разбиениями на случай корреляционной таблицы $k \times l$. В качестве исходной возьмем формулу, приводимую Б.Г. Миркиным и Л.Б. Черным в статье «Об измерении меры близости между различными разбиениями конечного множества объектов»[29].

Если R и S два разбиения множества из N элементов и R разбивает его на *m*, а S на *n* классов, причем в *i*-ом классе $|R_i|$ элементов, а в *j*-ом $|S_i|$ элементов, то мера близости разбиений

$$d(R,S) = \frac{1}{2}\sum_i |R_i|^2 + \frac{1}{2}\sum_j |S_j|^2 - \sum_i \sum_j |R_i \cap S_j|^2$$

(Здесь $R \cap S$ – пересечение классов R и S).

Так как $d_{\max} = \frac{1}{2}N(N-1)$, нормированная мера $\delta = \frac{2d}{N(N-1)}$, причем $0 \le \delta \le 1$, где $\delta = 0$ соответствует

[132]

[28] *Миркин Б.Г.* Новый подход к обработке социологической информации. – В кн.: Измерение и моделирование в социологии. Новосибирск, 1969.

[29] Автоматика и телемеханика, 1970, №5.

максимальной связи, δ = 1 минимальной (отсутствие связи).

Для корреляционной таблицы $\{N_{ij}\}$: признак X осуществляет разбиение общности N на k классов x_i, в каждом из которых $N(x_i)$ элементов: признак Y на l классов y_i, в каждом из которых $N(y_j)$ элементов.

Так как $N_{ij} = N(x_i) \cap N(y_j)$, то мера близости двух рассматриваемых разбиений

$$\delta(x,y) = \frac{1}{N(N-1)}\left[\sum_{i=1}^{k} N^2(x_i) + \sum_{j=1}^{l} N^2(y_j) - 2\sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}^2\right] \quad \text{(II,8,2)}$$

Допустим, что мы исследуем некоторое разбиение, осуществляемое Y, и хотим выяснить значимость ряда признаков $X^{(p)}$ (p = 1, 2, ...) для выявления данного разбиения.

Значимость $X^{(p)}$ будет тем большей, чем ближе разбиения, т.е. чем меньше δ ($X^{(p)}$, Y) $\equiv\delta_p$. Таким образом, значимость признака $X^{(p)}$ по отношению к разбиению Y можно принять обратно пропорциональной расстоянию δ_p. Эту значимость («силу влияния») можно интерпретировать, следуя Б.Г. Миркину, как меру связи между признаками. Пусть, например, разбиение У – это социально-профессиональные группы, а $X^{(p)}$ – различные социально-демографические признаки (профессия, квалификация, образование, доход, место жительства и т.д.), вычисляя δ_p, мы можем определить значимость (влияние) различных $X^{(p)}$ для выявления Y-разбиения, выделить наиболее информативные признаки.

Рассматривалась и такая задача: пусть Y – это расселение работников по «зонам доступности предприятия»[30], а X(p) – некоторые социально-демографические признаки, значимые для расселения. Наиболее значимым признаком оказалась принадлежность к социально-профессиональной группе.

Рассмотрим еще раз вопрос о связи между удовлетворенностями заработной платой и работой, используя для ее характеристики обсуждаемую меру (см. пример № 14 § 1 этой главы).

[133]

[30] «Зона доступности предприятия» определяется временем, затрачиваемым работником на передвижение от места жительства до места работы. По нормам градостроительства выделяются четыре зоны: А (до 30 мин.), Б (от 30 до 45 мин.), В (от 45 мин. до часа), Г (свыше часа).

Теперь $S = \dfrac{A = B - C}{D}$, где

$$A = \sum_{i=1}^{3} N^2(x_i) = 448^2 + 508^2 + 52^2 = 461472,$$

$$B = \sum_{j=1}^{3} N^2(y_j) = 682^2 + 97^2 + 229^2 = 526974,$$

$$C = 2\sum_i \sum_j N_{ij}^2 = 2(350^2 + 35^2 + 63^2 + 298^2 + 52^2 + 158^2 + 34^2 + 10^2 + 8^2) = 490972$$

$D = N(N–1)$=1015056, δ – 0,490.

Упражнение 69. Показать, что для таблицы 19 § 1 этой главы δ = 0,528.

В первом случае δ меньше, но так как связь пропорциональна 1/ δ то она больше, чем во втором; таким образом, сделанный ранее вывод (§ 1) подтверждается.

Δ-коэффициент (номинальные шкалы)

В работе И.А. Шкрабкиной и Г.И. Смирновой «Программа измерения тесноты связи между двумя признаками»[31] предлагается для измерения связи использовать модульный коэффициент Δ.

Наряду с корреляционной таблицей $\{N_{ij}\}$ рассмотрим таблицу $\{\tilde{n}_{ij}\}$, где $\tilde{n}_{ij} = \dfrac{N_{ij}}{N(x_i)}$ и введем величину $\bar{n}_j = \dfrac{1}{k}\sum_{p=1}^{k} \tilde{n}_{pj}$ Мерой связи, точнее, влияния X на Y, может служить

$$S = \sum_{i=1}^{k}\sum_{j=1}^{l} \left|\tilde{n}_{ij} - \bar{n}_j\right|$$

Покажем это. Если признаки независимы, то $\tilde{n}_{ij} = \dfrac{N(y_i)}{N}$, а $\bar{n}_j = \tilde{n}_{ij}$, т.е. рассматриваемая сумма обращается в нуль.

Логика измерения связи такова: если признаки незави-

[134]

[31] Анализ социологической информации с применением ЭВМ, ч.1. М., 1973, с.143-157

симы, то при изменении X значение Y не должно меняться, т.е. числа индивидов $\widetilde{n}_{ij}$ с разными X при фиксированном Y должны быть примерно одинаковы, т.е. равными $\overline{n}_j$. Если же X влияет на Y, то $\widetilde{n}_{ij}$ должны отличаться от среднего $\overline{n}_j$.

Обсуждаемая сумма не является нормированной. В работе Шкрабкиной и Смирновой в качестве коэффициента при сумме предлагается использовать величину $\frac{1}{k}$. Однако как легко видеть, $\Delta' = \frac{S}{k}$ не является нормированной величиной.

Для нормировки необходимо найти максимальное значение суммы S. Оказывается, что оно достигается в случае полной связи (связь мы называем полной, если каждому X соответствует одно значение Y) и равно $\frac{2}{k}(m-1)(2k-m)$, где m = min*(k,l)* – меньшее из чисел k и l.

Таким образом, нормированный коэффициент, описывающий влияние X на Y:

$$\Delta_{yx} = \frac{k}{2(m-1)(2k-m)} \sum_{i=1}^{k} \sum_{j=1}^{l} \left| \widetilde{n}_{ij} - \overline{n}_j \right| \qquad \text{(II,8,3)}$$

Итак, $0 \le \Delta \le 1$, причем 0 соответствует отсутствию, а 1 – полной связи.

Аналогично

$$\Delta_{xy} = \frac{k}{2(m-1)(2k-m)} \sum_{i=1}^{k} \sum_{j=1}^{l} \left| \widetilde{\widetilde{n}}_{ij} - \overline{\overline{n}} \right|,$$

где

$\widetilde{\widetilde{n}}_{ij} = \frac{N_{ij}}{N(y_j)}$, а $\overline{\overline{n}}_i = \frac{1}{l} \sum_{p=1}^{l} \widetilde{\widetilde{n}}_{ip}$.

Все ранее рассмотренные здесь коэффициенты применимы даже для номинальных шкал. Перейдем к коэффициентам, которые используются при наличии упорядочения значений признаков.

γ-коэффициент Гудмана

По определению

$$\gamma = \frac{P - Q}{P + Q} \qquad \text{(II,8,4)}$$

[135]

где P – число пар объектов, у которых оба признака упорядочены в одинаковой последовательности, а Q – то же, но в обратной.

Пусть значения X и Y в корреляционной таблице выписаны в одинаковой последовательности. Величину *P* можно вычислить как сумму результатов умножения частот каждой

Таблица 32

Пример расчета γ-коэффициента Гудмана

X	Y			$N(x_i)$
	y_1	y_2	y_3	
x_1	35	15	5	55
x_2	5	25	15	42
$N(y_j)$	40	40	20	100

клетки на сумму частот, расположенных в клетках ниже и правее:

$$P = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}\left(\sum_{r=i+1}^{k}\sum_{s=j+1}^{l} N_{rs}\right) . \quad \text{(II,8,5)}$$

Это выражение, очевидно, совпадает с уменьшаемым в формуле (II,6,9). Q – сумма результатов умножения частот каждой клетки на сумму частот, расположенных ниже и левее ее:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}\left(\sum_{r=i+1}^{k}\sum_{s=1}^{j-1} N_{rs}\right) \quad \text{(II,8,6)}$$

(Q – вычитаемое в упоминавшейся формуле).

Если связь полная и прямая, то Q = 0 и γ = 1, если же полная и обратная, то P = 0 и γ = – 1. Итак, $-1 \le \gamma \le 1$

Положительный γ-коэффициент Гудмана показывает, насколько вероятно, что при увеличении значения одного признака увеличится значение другого (отрицательный – при увеличении одного – уменьшается значение другого).

Так как этот коэффициент в наших социологических исследованиях еще не получил распространения, приведем пример его вычисления для простейшей таблицы 32.

P = 35(25+15)+15·15+5(25+15)+25·15=1625

$\tilde{Q}$ = 5(5+25)+15·5=225

γ = +0,76

[136]

Упражнение 70. Для таблицы 32 рассчитать γ -коэффициенты Гудмана. Ответ: 0,33; 0,44.

d-коэффициент Сомерса

По определению,

$$d_{yx} = \frac{P - Q}{P + Q + Y_0} \qquad \text{(II,8,7)}$$

$$d_{xy} = \frac{P - Q}{P + Q + X_0} \qquad \text{(II,8,8)}$$

где Y_0 – число пар объектов с одинаковыми значениями Y (но разными X), а X_0 – с одинаковыми X (но разными Y), P и Q определены выше, см. (II,8,5), (II,8,6).

Вообще говоря, $X_0 \neq У_0$ (далее мы рассмотрим способ их вычисления), следовательно, коэффициент *d* не является симметричным: $d_{yx} \neq d_{xy}$ Его следует применять, когда из содержательных соображений ясно, что влияние X на Y и Y на X неодинаково.

В частности, *d* используется, если не имеет смысла влияние, скажем, X на Y (удовлетворенность работой X не может влиять на возраст Y, хотя, например, может влиять на квалификацию). При этом вычисляется, естественно, лишь один коэффициент: в рассмотренном примере – d_{xy}, описывающий влияние Y на X.

Перейдем к вычислению X_0, т.е. числа пар объектов с одинаковыми X (но разными Y).

Для вычисления X_0 найдем сперва вклад $_i$-ой строки (все объекты этой строки имеют одинаковые значения X, равные x_i):

$N_{i1}N_{i2} \ldots N_{ij} \ldots N_{il}$

Число пар с одинаковыми X, но разными Y в этой строке:

$$N_{i1}(N_{i2} + N_{i3} + \ldots + N_{il}) + N_{i2}(N_{i3} + \ldots + N_{il}) + \ldots + N_{il-1}N_{il} = \sum_{p=1}^{l-1} N_{ip} \sum_{q=p+1}^{l} N_{iq}$$

Вклад всех строк и составляет X_0:

$$X_0 = \sum_{i=1}^{k} \sum_{p=1}^{l-1} N_{ip} \sum_{q=p+1}^{l} N_i$$

[137]

Аналогично:

$$Y_0 = \sum_{j=1}^{l} \sum_{p=1}^{k-l} N_{pj} \sum_{q=p+1}^{k} N_{qj}$$

Замечание. Так как число пар с одинаковыми X и Y

$$Z_0 = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{l} N_{ij} (N_{ij} - 1),$$

то

$$Z_0 + Y_0 + X_0 + P + Q = \frac{N(N-1)}{2}$$

Это соотношение можно применять для контроля вычислений. Для таблицы 32 вычислим коэффициенты Сомерса:

$X_0 = 35 \cdot (15 + 5) + 15 \cdot 5 + 5\ (25 + 15) + 25 \cdot 15 = 1350;$

$У_0 = 35 \cdot 5 + 15 \cdot 25 + 5 \cdot 15 = 625;$

$d_{yx} = +0{,}57$

$d_{xy} = +\ 0{,}53$

Близость полученных значений d_{xy} и d_{yx} можно интерпретировать как «симметрию» влияния X на Y и Y на X. Легко видеть, что $|d| \leq 1$ во всех случаях, причем $d = 0$, если связи нет. Приведем один пример использования рассмотренных коэффициентов в прикладных исследованиях.

Коэффициент γ широко применялся эстонскими социологами Института истории АН ЭССР при изучении удовлетворенности трудовой деятельностью. Согласно данным Т. Китвеля, ранжировка по γ оценок различных элементов рабочей ситуации по степени их связи с удовлетворенностью работой на данном предприятии имеет следующий вид: 1) содержание труда (0,597); 2) заработная плата (0,365); 3) сплоченность коллектива (0,340).

Далее идут: отношения с администрацией, организация труда и т.д.[32]

Обратим внимание на то, что эта последовательность сходна с той, которая была получена ленинградскими («Человек и его работа») и немецкими[33] социологами, а также находится в согласии с нашими результатами.

В наших исследованиях использовались: коэффициент Чупрова *T*, вариационный размах оценок, энтропийная мера связи λ. Все три способа дали одну и ту же последова-

[138]

[32] Китвель Т.О социально-психологических проблемах удовлетворенности трудом. Таллин, 1974, с. 75.

[33] Stollberg R. Arbeitszufriedenheit – theoretische und praktische probleme. Berlin, 1967, S.49

тельность элементов: содержание труда, организация труда, заработная плата, отношения с администрацией и т.д. Отметим, что указанную последовательность элементов мы получили как с помощью показателя двусторонней связи – коэффициента Чупрова, так и с помощью показателя односторонней (направленной) связи – энтропийной меры связи.

Использованный Китвелем коэффициент γ является мерой двусторонней связи. Представляется целесообразным также применение несимметричного коэффициента Сомерса, который, учитывает последовательность позиций на шкале удовлетворенности (в этом его несомненное преимущество перед и *T*, и λ) и является «направленным» (в отличие от *T* и γ). С его помощью можно описать влияние частных удовлетворенностей (т.е. различными элементами) на интегральную удовлетворенность работой.

Существуют также некоторые коэффициенты, которые разработаны для случаев, когда одна переменная измерена по номинальной, а вторая – порядковой или метрической шкале. Мы рассмотрим два из них.

Ранговый бисериалъный коэффициент[34]

Предназначен для случая, когда одна шкала номинальная дихотомическая, а вторая – порядковая. Его название связано с тем, что при этом есть как бы две серии данных: каждая серия для одного из значений дихотомической переменной.

Назовем ранговым бисериальным следующий коэффициент (формула пригодна при отсутствии объединенных рангов):

$$r_{\rho b} = \frac{2}{N}(\bar{y}_1 - \bar{y}_2) \quad \text{(II,8,9)}$$

где N – число объектов; $\bar{y}_1$ – средний ранг по признаку Y объектов, имеющих значение x_1 дихотомической переменной X; $\bar{y}_2$ – средний ранг объектов, имеющих значение x_2. Пусть, например, дана дихотомическая переменная X ($x_1 = 1$, $x_2 = 2$) и ранговая переменная Y:

признак X	1	2	1	2	1	1	2	2	1	1
признак Y	1	10	2	9	5	8	4	7	3	6

В первой строке стоят значения признака X, а во второй – ранги признака Y для некоторых 10 объектов. Выпишем ранги по Y для каждого значения признака X:

[139]

X	Y	
x_1=1	1, 2, 5, 8, 3, 6	$\bar{y}_1 = \frac{25}{6} = 4{,}167$
x_2=2	10, 9, 4, 7	$\bar{y}_2 = \frac{30}{4} = 7{,}500$

Точечно-бисериальный коэффициент корреляции[35]

Предназначен для изучения связи признаков, один из которых измерен в номинальной дихотомической, второй – в метрической шкале:

$$r_{rb} = \frac{\bar{y}_1 - \bar{y}^2}{\sigma_y N}\sqrt{\frac{(N-1)N(x_1)N(x_2)}{N}} \quad \text{(II,8,10)}$$

где $\bar{y}_1$ – среднее значение признака Y для объектов, имеющих значение x_1 а $\bar{y}_2$ – значение x_2 дихотомической переменной X; N (x_1) и N (x_2) – число объектов, имеющих значение x_1 и x_2 соответственно, N – число всех объектов, σ_y – среднее квадратическое

[34] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М., 1976, с. 165 – 167.

[35] Там же, с. 149-151. Отметим, что в таблице на с. 151 этой книги, видимо, опечатка в данных о росте, поэтому приведенные в ней результаты неверны.

отклонение для всех объектов. Аналогично предыдущему коэффициенту рассмотрим следующую таблицу:

Значения X	Значения Y
x_1=1	170; 140; 157; 152; 155; 160; 152
x_2=2	150; 160; 165; 183; 163; 168; 160; 157

$N(x_1)=7$, $N(x_2)=8$, $N=15$, $\bar{y}_1=155{,}14$, $\bar{y}_2=163{,}25$, $\sigma_y=9{,}31$, $r_{rb}=0{,}42$

Формула (II,7,10) представляет собой алгебраическое упрощение коэффициента *r* для случая, когда X – дихотомическая переменная, поэтому все расчеты можно было бы проводить и по формулам для *r*, например, (II,5,1) или (II,5,3). Обобщения этих коэффициентов (полисериальные коэффициенты) не получили широкого распространения.

[140]

Глава III

РЕГРЕССИИ

1. Основные понятия. Прямая регрессия. Криволинейные связи. Корреляционное отношение

Как отмечалось, при исследовании связи между двумя признаками находят распределение совокупности в виде корреляционной таблицы $\{N_{ij}\}$; тесноту связи характеризуют с помощью коэффициентов корреляции (глава II), а форму – с помощью уравнений регрессии, к рассмотрению которых мы и переходим.

Напомним, что каждому значению x_i, соответствует распределение *y*: y_j , N_{ij}, где $j=\overline{1,l}$. Такие распределения называют условными, условными называют и соответствующие средние

$$\overline{y_i}=\frac{\sum_{j=1}^{l} y_j N_{ij}}{N(x_i)},(i=\overline{1,k}) \qquad \text{(III,1,1)}$$

Полную среднюю $\bar{y}$ можно рассматривать как взвешенную сумму условных средних $\overline{y_i}$.

Упражнение 71. Показать, что $\bar{y}$, равное, по определению, $\frac{1}{N}\sum_{j=1}^{l} y_i N(y_j)$ равно $\frac{1}{N}\sum_{i=1}^{k}\overline{y_i}N(x_i)$.

Далее мы будем изучать связь $\overline{y_i}$, с x_i. Если ее можно представить в виде $\bar{y}=f(x_i)$, где *f(x)* – некоторая известная функция, то уравнение $\overline{y_i}=f(x)$, следуя Гальтону, называют *уравнением регрессии Y на X*, а соответствующую ему кривую – *кривой регрессии*[1]. С таким уравнением мы уже встречались в примере 42 (§1 главы II).

[141]

[1] Индекс *x* показывает, что речь идет об условном среднем.

Аналогично (III,1,1) определяется условная средняя

$$\overline{x_i} = \frac{\sum_{i=1}^{i} x_i N_{ij}}{N(y_j)}, \qquad \text{(III,1,2)}$$

соответствующая y_j (III, 1,2).

Упражнение 72. Показать, что $\overline{x}$ является взвешенной суммой условных средних $\overline{x_j}$; т.е. что

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{i} N(y_j)\overline{x_j}, \qquad \text{(III,1,3)}$$

Уравнение $\overline{x_y} = \varphi(y)$ называется уравнением регрессии X на Y. Подчеркнем, что, вообще говоря, обе регрессии – Y на X и X на Y – различны; влияния X на Y и Y на X не одинаковы. Следовательно, функции f и φ не являются взаимно обратными.

Пример 27. В ряде случаев связь удается представить в виде линейной зависимости типа $\overline{y_x} = ax + b$ и соответственно $\overline{x_y} = cy + d$.

Рассмотрим такую корреляционную таблицу для признаков X и Y.

X	Y					$N(x_i)$	$\overline{y_x}$
	y_1 =20	y_2 =30	y_3=40	y_4=50	y_5=60		
x_1=10	38	37	42	0	0	117	30,3
x_2=20	0	47	40	48	0	135	40,1
x_3=30	0	0	41	28	39	108	49,8
$N(y_j)$	38	84	123	76	39	360	---
$\overline{x_y}$	10,0	15,6	19,9	23,7	30,0	---	---

Упражнение 73. Вычислить $\overline{y_x}$ и $\overline{x_y}$ по данным таблицы примера 27 (ответы выписаны в соответствующей колонке и строке этой таблицы). Исходя из значений $\overline{y_x}$ и $\overline{x_y}$, приведенных в крайних маргиналах, можно записать приближенные равенства:

$$\overline{y_x} = x + 20 \qquad \text{(III,1,4)}$$

$$\overline{x_y} = 0{,}5 * y \qquad \text{(III,1,5)}$$

[142]

В дальнейшем мы рассмотрим нахождение уточненных уравнений регрессии, а сейчас подчеркнем, что уравнения (III 1,4) и (III,1,5) существенно различны: из одного нельзя получить другое. В этом, в частности проявляется специфика корреляционных связей. Иное дело – связи функциональные. Получаемые для опытных данных регрессии $\bar{y}_x = f(x)$ и $\bar{x}_y = \varphi(y)$, являющиеся выражением одной и той же функциональной связи, должны быть в случае надежных данных взаимно обратными. (Кстати, взаимная обратность функций *f* и φ является обычно критерием надежности эмпирического материала).

Наша задача заключается в нахождении уравнения регрессии. Как она решается, рассмотрим на примере прямой регрессии общего вида, а затем вернемся к нашему примеру.

Прямая регрессия

О прямой (точнее – прямолинейной) регрессии говорят в том случае, когда точки $(x_i, \bar{y}_i)$ располагаются близко к некоторой прямой $y=ax+b$. Уравнение регрессии будет полностью известно, если мы найдем *a* и *b*. Естественным условием их нахождения является минимум отклонений эмпирических точек $(x_i, \bar{y}_i)$ от прямой, являющейся линией регрессии.

Мерой отклонения опытных точек от прямой может служить величина дисперсии

$$S = \frac{1}{N}\sum_{i=1}^{k} N(x_i)(\bar{y}_i - y_i)^2, \qquad \text{(III,1,6)}$$

где $y_i = ax_i + b$ – теоретическое значение *Y*, соответствующее x_i, а $\bar{y}_i$ –эмпирическое среднее, определяемое соотношением (III,1,1).

В *S*-отклонение $\bar{y}_i$ от y_i входит: 1) в квадрате, так как не должны компенсироваться отклонения разных знаков; 2) со своим «удельным весом» $\frac{N(x_i)}{N}$.

У нас $S=S(a, b)$. Параметры *a* и *b* найдем из условия минимума *S*, т.е. суммы квадратов отклонений (отсюда и название способа – «метод наименьших квадратов»).

Представим уравнение регрессии $y=ax+b$ в виде

$$y - \bar{y} = a(x - \bar{x}) + c, \qquad \text{(III,1,7)}$$

где $c = b - \bar{y} + a\bar{x}$.

[143]

Теперь

$$S = S(a,c) = \frac{1}{N}\sum_{i=1}^{k} N(x_i)\left[\overline{y}_i - \overline{y} - a(x_i - \overline{x}) - c\right]^2, \qquad \text{(III,1,8)}$$

и задача свелась к нахождению a и c, обеспечивающих минимум S.

S можно рассматривать как взвешенную сумму квадратов отклонений величины $\overline{y}_i - \overline{y} - a(x_i - \overline{x})$ от c. Согласно четвертому свойству дисперсии (§3 главы 1) S достигает минимума, когда c равно среднему значению величины $\overline{y}_i - \overline{y} - a(\overline{x}_i - x)$, т.е.

$$c = \frac{1}{N}\sum_{i=1}^{k} N(x_i)[\overline{y}_i - \overline{y} - a(x_i - \overline{x})] = 0.$$

Здесь мы использовали соотношения (III,1,1) и определения $\overline{x}$ и $\overline{y}$, Величину a нужно найти из условия минимума

$$S(a,0) = \frac{1}{N}\sum N(x_i)[\overline{y}_i - \overline{y} - a(x_i - \overline{x})]^2 \qquad \text{(III,1,9)}$$

Читатель, знакомый с элементами высшей математики, легко поймет, что условие минимума $\frac{\partial S}{\partial a} = 0$ принимает вид

$$\sum N(x_i)[\overline{y}_i - \overline{y} - a(x_i - \overline{x})](x_i - \overline{x}) = 0 \qquad \text{(III,1,10)}$$

Откуда

$$a = \frac{\sum N(x_i)(\overline{y}_i - \overline{y})(x_i - \overline{x})}{\sum N(x_i)(x_i - \overline{x})^2} \qquad \text{(III,1,11)}$$

Упражнение 74, Убедиться, что при этом $\frac{\partial^2 S}{\partial a^2} > 0$, т.е. действительно имеет место минимум.

Для читателя, не знакомого с высшей математикой, заметим, что a можно найти также с помощью соображений, основанных на элементарной математике.

Действительно, перепишем S в виде

$$[\frac{1}{N}\sum N(x_i)(x_i - \overline{x})^2]a^2 - 2[\frac{1}{N}\sum N(x_i)(\overline{y}_i - \overline{y})(x_i - \overline{x})]a +$$

$$+\frac{1}{N}\sum N(x_i)(\overline{y}_i - \overline{y})^2 = Aa^2 - 2Ba + B =$$

$$= A\left(a - \frac{B}{A}\right)^2 B - \frac{B^2}{A}$$

[144]

(где смысл обозначений A, B, D очевиден). Минимальное значение S, равное $D-\frac{B^2}{A}$, достигается при

$$a=\frac{B}{A}=\frac{\sum N(x_i)(\overline{y_i}-\overline{y})(x_i-\overline{x})}{\sum N(x_i)(x_i-\overline{x})^2}.$$

Тем самым мы независимо обосновали справидливость (III,1,11).

Это обстоятельство будет использовано в дальнейшем.

Зная a, из условия c=0 можно найти b:

$$b=\overline{y}-a\overline{x}. \qquad \text{(III,1,12)}$$

Тем самым полностью определено уравнение регрессии. Обратим внимание на то, что мы здесь фактически доопределили, уточнили понятие уравнения регрессии. Раньше таковым называлось уравнение $\overline{y_i}=f(x_i)$ (в случае регрессии Y на X). Теперь мы видим, что уравнение регрессии описывает кривую, отклонение эмпирических точек ($x_i, \overline{y_i}$) от которой минимально. Ясно, что задача отыскания «точной» кривой, на которой лежат эти точки, и очень сложна и нецелесообразна. Доопределенное уравнение регрессии, способ нахождения которого здесь рассмотрен, позволяет сравнительно просто и надежно судить о форме связи между переменными.

Упражнение 75. По данным корреляционной таблицы примера 27 найти уравнения регрессии Y на X и X на Y.

Указание. Использовать формулы (III,1, 11), (III,1,12).

Ответ: $\overline{y_x}=0{,}974\cdot x+20{,}58$ (III,1,4')

$\overline{x_y}=0{,}468\cdot y+1{,}11$ (III,1,5')

Эти соотношения являются уточнением уравнений (III,1,4), (III,1,5), которые были получены «на глазок».

Теперь уравнение регрессии (III,1,7) принимает вид:

$$y-\overline{y}=r\frac{\sigma_y}{\sigma_x}(x-\overline{x}), \qquad \text{(III,1,13)}$$

где

$$r=\frac{1}{N\sigma_x\sigma_y}\sum_i N(x_i)(\overline{y_i}-\overline{y})(x_i-\overline{x}).$$

Упражнение 76. Показать, что:

1. $$r=\frac{1}{N\sigma_x\sigma_y}\sum_i\sum_j N_{ij}(\overline{y_j}-\overline{y})(x_i-\overline{x}) \qquad \text{(III,1,14)}$$

2. $r = \frac{1}{N\sigma_x\sigma_y}(\overline{xy} - \bar{x}\cdot\bar{y})$ (III,1,15)

3. $r = \frac{1}{N\sigma_x\sigma_y}\sum_j N(y_i)(y_j - \bar{y})(\overline{x_i} - \bar{x}).$ (III,1,16)

Рассмотрим, например,

$$\sum_{i,j} N_{ij}(y_j - \bar{y})(x_i - \bar{x}) = \sum_i (x_i - \bar{x})\sum_j N_{ij}(y_j - \bar{y}) =$$

$$= \sum_i N(x_i)(x_i - \bar{x})(\overline{y_i} - \bar{y}).$$

Таким образом, мы показали, что в уравнение регрессии входит ранее определенная величина r (§ 5 главы II) и тем самым пришли к парному коэффициенту корреляции из теоретических соображений.

Наиболее простую интерпретацию r допускает в так называемых нормальных координатах. Введем $t_x = \frac{x - \bar{x}}{\sigma_x}$ и $t_y = \frac{y - \bar{y}}{\sigma_y}$. Новые переменные безразмерны, имеют нулевые средние и единичные дисперсии. Они не зависят от масштаба.

В этих переменных уравнение регрессии принимает вид:

$t_y = rt_x$. (III,1,17)

Таким образом, r показывает, на сколько изменяется зависимая переменная при изменении независимой на единицу. Величина $\rho_{yx} = r\frac{\sigma_y}{\sigma_x}$ угловой коэффициент уравнения регрессии Y на X.

Упражнение 77. Показать, что регрессия X на Y имеет вид:

$$x - \bar{x} = r\frac{\sigma_y}{\sigma_x}(y - \bar{y}).$$ (III,1,18)

Теперь $\rho_{xy} = r\frac{\sigma_x}{\sigma_y}$. Ясно, что произведение угловых коэффициентов в уравнениях регрессии Y на X и X на Y равно квадрату коэффициента парной корреляции, а регрессии совпадают только в том случае, когда $|r| = 1$.

При подстановке в уравнение регрессии координат точек $(x_i, \overline{y_i})$ мы получим точное равенство только в том случае, когда все эмпирические точки лежат на одной прямой. На практике этого не бывает и равенство $\overline{y_i} - y = \rho_{yx}(x_i -$

[146]

$-\overline{x})$ выполняется приближенно. В качестве меры точности естественно принять среднеквадратическую погрешность, т.е. квадратный корень из отклонения (дисперсии). Мера точности, таким образом $\sqrt{S_{\min}}$, где

$$S_{min} = D - \frac{B^2}{A} = \frac{1}{N}\sum N(x_i)(\overline{y}_i - \overline{y})^2 - \\ - \frac{\left[\sum N(x_i)(x_i - \overline{x})(\overline{y}_i - \overline{y})\right]^2}{N\sum N(x_i)(x_i - \overline{x})^2}, \qquad \text{(III,1,19)}$$

если учесть определения A, B и D.

До сих пор мы рассматривали прямолинейную регрессию, используя метод наименьших квадратов. Этот метод может быть применен и для изучения криволинейной зависимости. В некоторых случаях не потребуется решать криволинейную задачу, ее можно свести к рассмотренной. Для этого используется замена переменных.

Мы приведем интересный социально-демографический пример в форме своеобразного *упражнения* (*№78*): часть выкладок читателю предстоит выполнить самостоятельно. (Впрочем, понять смысл рассматриваемого примера можно и не прибегая к несколько громоздким, хотя и несложным выкладкам, которые предлагаются читателю по ходу изложения материала).

Пример 28. В 1965 г. И. С. Шкловским был установлен гиперболический закон роста численности населения земного шара на материале статистики с 1600 г. по 1960 г. Математически он выглядит так: $y(x) = \frac{A}{B - x}$, где x – календарное время, $y(x)$ – численность населения, а A и B – параметры уравнения. Статистический материал, которым располагал Шкловский[2], приведен в табл. 33.

Сделаем замену переменных: перейдем от x к $X'=x-x_0$,, где x_0 – начало отсчета времени, т.е. 1600, и от y к $Y' = \frac{1}{y}$.

Построив график Y'=Y'(X'), видим, что все точки тесно группируются возле прямой линии. (Убедитесь самостоятельно. Именно здесь начинается для читателя само-упражнение. Кстати, постройте график y=y(x), убедитесь, что точки ложатся на гиперболу).

[147]

[2] Таблица заимствуется из книги: *Суслов И. П., ГражданниковЕ.*Д. Основы социальной статистики, Новосибирск, 1973 (мы несколько уточнили приведенные авторами расчеты и устранили имеющиеся опечатки).

В силу сказанного, станем искать *Y′(X′)* в виде $-aX'+b$, используя метод наименьших квадратов. (Знак минус показывает, что с ростом *X′* величина *Y′* уменьшается – так и должно быть: ведь $Y'=\frac{1}{y}$).

Таблица 33

Численность населения земного шара

Год	Численность (млн. чел)	Рассчитанная численность (млн. чел)	Отклонения (%)
1600	486	481	1,0
1650	545	545	0
1700	617	627	-1,7
1750	728	739	-1,6
1800	906	900	0,6
1850	1171	1150	1,8
1900	1608	1592	1,0
1910	-	1725	-
1920	1861	1882	-1,1
1930	2070	2070	0
1940	2295	2300	0,2
1950	2517	2588	-2,8
1960	3010	2958	1,7

Теперь

$$a = \frac{\sum X_i' \sum Y_i' - N \sum X_i' Y_i'}{N \sum (X_i')^2 - (\sum X_i')^2}$$

$$b = \frac{\sum (X_i')^2 \sum Y_i' - \sum X_i' \sum X_i Y_i'}{N \sum (X_i')^2 - (\sum X_i')^2}$$

Это несложно показать, если внимательно рассмотреть материал данного параграфа. Для каждого x_i, можно вычислить x'_i и y'_i, и, следовательно, найти *a* и *b* (сделайте это),

Теперь $y(x)=\frac{A}{B-x}$, где $A=\frac{1}{a}$, $B=x_0+\frac{b}{a}$

После соответствующих вычислений получим: A = 207052, B = 2030, т.е. окончательно:

$y(x)=\frac{207052}{2030-x}$ — закон Шкловского.

Найдем расчетную численность. Эти данные приводятся в таблице (колонка 3). Подсчет относительных отклонений

[148]

показывает, что они не превосходят по абсолютной величине 2,8 (колонка 4).

Итак, получено теоретическое уравнение. Читатель вправе задать вопрос: «Ну и что? Для чего это уравнение? Что оно дает нам? Значения y_i, которые были известны заранее, да и то, как видно из таблицы, приближенно?!»

Попытаемся ответить. Мы установили закономерность, которой подчиняется эмпирический материал, а знание закономерности может стать источником новых сведений. Но экстраполируя данные, полученные с помощью формулы, на прошлое и будущее, нужно помнить, что наши предсказания будут тем надежней, чем меньше выбираемый интервал. Например, из формулы Шкловского следует, что к 2030 г. население должно стать бесконечно большим. Этот результат, конечно, не имеет, как принято говорить, «физического» смысла, что отнюдь не свидетельствует о неправильности формулы. Просто нужно помнить, что обычно закономерности относятся ко вполне определенным условиям, что устанавливаемые формулы имеют границы применимости. Так, мы с достоверностью не можем, зная закон Шкловского, вычислить величину народонаселения, скажем, в 1500 или 2000 году. Расчеты для 1970 и 1980 годов по этой формуле дают 3450 и 4140 млн. человек, что на 5,1 и 6,3% ниже реальной численности (3635 и 4415 млн. соответственно). Хотя ошибка несколько возрастает, формула дает очень хорошее приближение к реальным данным.

Можно предположить, что в ближайшие десятилетия мы станем свидетелями изменения темпов роста населения земного шара – закон перестанет быть гиперболическим. Это, само собой, нисколько не опровергает формулу Шкловского, установленную для рассмотренных временных интервалов. Отметим, что она дает возможность определять численность населения в те годы внутри изученного интервала, для которых статистика отсутствует или ненадежна. Так, в 1910 г. население примерно составляло 1725 млн. человек и т.д.

Корреляционное отношение

Вернемся, однако, к рассмотрению регрессий. В случае *криволинейной* зависимости целесообразно использовать так называемое корреляционное отношение

$$\eta_{нч} = \frac{\sqrt{\frac{1}{N}\sum N(x_i)(\overline{y_i} - \overline{y})^2}}{\sigma_y}, \qquad \text{(III,1,20)}$$

[149]

которое, по определению, представляет собой отношение среднего квадратического отклонения условных средних $\overline{y_i}(\sigma_{\bar{y}})$ к полному среднему квадратическому отклонению (σ_y): $\eta_{yx} = \sigma_{\bar{y}} / \sigma_y$ (см. § 5 главы II).

С учетом (III,1,13), (I1I,1,19), (III,1,20)

$min S = \sigma_y^2 (\eta_{yx}^2 - r^2)$.

Так как, по определению, $S \geq 0$, то $\eta_{yx}^2 \geq r^2$ или $\eta_{yx} \geq |r|$. Итак, *minS=0,* если все (x_i, $\overline{y_i}$) лежат на одной прямой, т.е. регрессия *Y* на *X* прямолинейная. Таким образом, равенство является условием того, что регрессия прямолинейная. Во всех остальных случаях (криволинейная зависимость!)

$\eta_{yx} > |r|$.

Мы видели (§ 4 главы III), что $0 \leq \eta \leq 1$. Можно аналогично показать, что $-1 \leq r \leq 1$. Доказательство справедливости этого утверждения составит содержание следующего *упражнения.*

Упражнение 79.

Указание. Использовать очевидное неравенство

$$\sum_{i,j} N_{ij} \left[y_i - \bar{y} - r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) \right]^2 \geq 0,$$

преобразуя его к виду $\sigma_y^2 (1 - r^2)$. Тогда $1 - r^2 \geq 0$, т.е. $|r| \leq 1$.

Итак, мы нашли диапазон возможных значений, принимаемых *r* и η, выяснили условие того, что регрессия прямолинейная и нашли меру криволинейной связи (η). Так как обычно связи криволинейные, следует обратить особое внимание на корреляционное отношение.

К сожалению, в социологической литературе, как уже отмечалось, наблюдается злоупотребление коэффициентом *r*, который вычисляется без обоснования правомерности его использования. Лишь в редких случаях исследователи применяют η, хотя ситуация должна быть обратной.

Упражнение 80. Показать, что в случае корреляционной таблицы:

$$\eta_{yx}^2 = \frac{N \sum \frac{1}{N(x_i)} (\sum N_{ij} y_j)^2 - \left[\frac{1}{N} \sum N(y_i) y_i \right]^2}{N \sum N(y_j) y_j^2 - \left[N^{-1} \sum N(y_j) y_j \right]^2} \qquad \text{(III,1,21)}$$

[150]

Вернемся к рассмотрению $\eta_{yx} = \sigma_{\bar{y}} / \sigma_y$. Стоящая в числителе величина $\sigma_{\bar{y}}$ описывает колеблемость *Y* под влиянием *X*. σ_y описывает полную колеблемость величины *Y* под влиянием всех условий. Следовательно, η_{yx} показывает, какую часть общей изменчивости *Y* обусловливает влияние *X*. Это отношение выявляет степень воздействия *X* на *Y*.

Таблица 34

Пример расчета корреляционного отношения

Возраст, лет (*X*)	Выполнение нормы выработки, % (*Y*)					$N(x_i)$
	95-100	100-105	105-110	110-115	115-120	
19-22	5	7	2	4	4	22
22-25	1	7	2	3	12	21
25-28	3	2	2	8	13	28
28-31	1	1	3	1	5	11
31-34	0	0	3	5	3	11
34-37	0	0	1	2	5	8
37-40	0	0	3	2	4	9
40-43	0	0	0	0	0	0
43-46	0	0	0	0	1	1
46-49	0	0	0	0	0	0
49-52	0	1	0	0	0	1
$N(y_j)$	10	14	16	25	47	112

Аналогично η_{yx} может быть определена величина η_{xy} , которая характеризует воздействие Y на X:

$$\eta_{xy} = \frac{\sigma_{\bar{x}}}{\sigma_x} \qquad \text{(III,1,22)}$$

Вообще говоря $\eta_{xy} \neq \eta_{yx}$, ибо воздействия *X* на *Y* и *Y* на *X* неравнозначны. Поэтому целесообразно вычислять оба корреляционных отношения, если они имеют содержательный смысл. Для *Y* – производительности труда рабочих, а *X* – стажа значение η_{yx} можно рассматривать как степень влияния стажа на производительность, корреляционное отношение η_{xy} в данном случае интерпретировать нельзя.

Упражнение 81. Записать выражение для η_{xy}.

Упражнение 82. Для таблицы 34 рассчитать корреляционное отношение[3]. Указание: Для вычислений удобно

[151]

[3] Данные заимствованы из «Методики и техники...», с.150.

перейти к $x' = \frac{x-a}{\alpha_x}$ и $y' = \frac{x-b}{\alpha_y}$, полагая a=35,5; b=107,5; α_x=3, α_y=5 (убедиться, что η при этом не изменится!)

В новых переменных x'_i, y'_j корреляционное отношение

$$\eta^2_{yx} = \frac{\sum_{i=1}^{11}[N(x_i)]^{-1}(\sum_{j=1}^{5} y'_j N_{ij})^2 - \frac{1}{N}\left[\sum_{j=1}^{5} y'_j N(y_j)\right]^2}{\sum_{j=1}^{5} y'^2_j N(y_j) - \frac{1}{N}\left[\sum_{j=1}^{5} y'_j N(y_j)\right]^2}$$

С учетом данных таблицы имеем:

η_{yx}=0,41.

(Читатель, испытывающий затруднения при вычислении этого коэффициента, может обратиться к с.150 – 151 «Методики и техники…», где найдет подробные выкладки.

Упражнение 83. По данным последней таблицы рассчитать *r*. Для этой цели удобно использовать формулу (11,5,4). Ответ: 0,21.

Итак, $r<\eta$. Связь нелинейная[4]. Для установления ее формы целесообразно построить эмпирическую кривую регрессии по точкам $(x_i, \overline{y_i})$. Эта работа составит содержание *упражнения 84*.

2. Частная корреляция. Случай трех признаков

Наличие статистической связи между двумя величинами может быть следствием связи обеих с некоторой третьей (либо совокупностью некоторых величин). Следовательно, возникает необходимость устранить влияние «третьих» величин. Заметим, что в простейшем случае этого можно достичь, изучая связи между двумя данными величинами в совокупности однородных объектов (при фиксированном «третьем» признаке). Однако для такой процедуры необходимы большие общности, особенно если устраняется влияние не одного, а нескольких признаков. Для изучения связи в таких ситуациях служит специальный аппарат частной корреляции. Рассмотрим принципиальную схему метода. Если корреляция данных признаков уменьшается при устранении неко-

[152]

[4] Значимость отклонения от линейности определяется с помощью критерия Фишера *(Закс* Л. Статистическое оценивание. М., 1976, с. 401).

торого признака, то взаимозависимость выделенных признаков определяется, в частности, и этим признаком. В предельном случае, когда устранение обращает коэффициент корреляции в нуль, можно считать, что этот признак обусловливает изучаемую связь.

Если при устранении коэффициент корреляции увеличивается, то данный признак ослабляет связь. Если же коэффициент корреляции практически не меняется, то соответствующий признак на связь не влияет.

Рассмотрим одну содержательную задачу. При изучении связи между производительностью труда и возрастом было установлено наличие прямой корреляции. Но на производительность влияет и стаж работы, который тоже оказывается в прямой корреляции с возрастом и с производительностью. Чтобы выяснить, прямая или обратная связь между производительностью и собственно возрастом, нужно, очевидно, устранить влияние стажа. Решить этот вопрос, непосредственно сопоставляя между собой три полученных парных коэффициента корреляции, невозможно. (Забегая вперед, укажем, что связь между производительностью и возрастом при устранении стажа оказалась отрицательной, а между производительностью и стажем при устранении возраста положительной, но более тесной).

Перейдем к рассмотрению техники частной корреляции, ограничившись для простоты выкладок случаем трех признаков. (Рассмотрение общего случая не потребует новых идей, хотя и оказывается значительно более громоздким).

Допустим, что изучаемая совокупность из N объектов может быть описана с помощью количественных признаков Y, X_1, и X_2. (Во избежание недоразумений подчеркнем, что теперь X_i, – не i-ое значение признака, как было раньше, а сам i-ый признак (i=1, 2), который может, в свою очередь, принимать ряд различных значений). Если признак Y принимает m различных значений y_g ($g = \overline{1,m}$), то $\bar{y} = \frac{1}{N}\sum_{g=1}^{m} N_g y_g$, где N_g – число индивидов, у которых $Y=y_g$. Обозначим через $\overline{x_{ig}}$ среднее значение признака X_i, у индивидов с $Y=y_g$, тогда $\overline{x_i} = \frac{1}{N}\sum_{g=1}^{m} N_g \overline{x_{ig}}$.

Если, например, Y – квалификация, а X_1 – возраст рабочих некоторого коллектива из N индивидов, то $y_g = g$ ($g = \overline{1,6}$) – тарифно-квалификационный разряд (для

[153]

определенности, предполагается, что сетка имеет 6 разрядов). N_g – число рабочих, у которых разряд g, $\overline{x_{1g}}$ – средний возраст рабочих с разрядом g, а $\overline{x_1}$ – средний возраст рабочих данного коллектива. Аналогично интерпретируется $\overline{x_{2g}}, \overline{x_2}$, если X_2, скажем, стаж работы и т.д. Найдем линейную зависимость Y от X_i (i=1, 2), которая удовлетворяет принципу наименьших квадратов. Для этого введем величину $\delta y = y - \overline{y}$ и $\delta x = x - \overline{x}$. Теперь указанная выше зависимость, по аналогии с предыдущим, может быть представлена в виде $\delta y = a_1 \delta x_1 + a_2 \delta x_2$.

Найдем a_i (i=1, 2) из условия минимума суммы квадратов отклонений.

$$S = S(a_1 a_2) = \sum_g N_g (\delta y_g - a_1 \delta x_{1g} - a_2 \delta x_{2g})^2,$$

условия минимума по аналогии с (III,1,10) принимают вид:

$$\begin{cases} \sum N_g (\delta y_g - a_1 \delta x_{1g} - a_2 \delta x_{2g}) \delta x_{1g} = 0 \\ \sum N_g (\delta y_g - a_1 \delta x_{1g} - a_2 \delta x_{2g}) \delta x_{2g} = 0 \end{cases}$$

Или:

$$\begin{cases} a_1 \sum N_g \delta x_{1g}^2 + a_2 \sum N_g \delta x_{1g} \delta x_{2g} = \sum N_g \delta x_{1g} \delta y_g \\ a_1 \sum N_g \delta x_{1g} \delta x_{2g} + a_2 \sum N_g \delta x_{2g}^2 = \sum N_g \delta x_{2g} \delta y_g \end{cases}$$

С использованием определения a получаем:

$$\begin{cases} a_1 \sigma_1^2 + a_2 \sigma_1 \sigma_2 \dfrac{\sum N_g \delta x_{1g} \delta x_{2g}}{N \sigma_1 \sigma_2} = \sigma_0 \sigma_1 \dfrac{\sum N_g \delta x_{1g} \delta y_g}{N \sigma_0 \sigma_1} \\ a_1 \sigma_1 \sigma_2 \dfrac{\sum N_g \delta x_{2g} \delta x_{1g}}{N \sigma_1 \sigma_2} + a_2 \sigma_2^2 = \sigma_0 \sigma_2 \dfrac{\sum N_g \delta x_{2g} \delta y_g}{N \sigma_0 \sigma_2} \end{cases}$$

Но $\dfrac{\sum N_g \delta x_{1g} \delta x_{2g}}{N \sigma_1 \sigma_2} = r_{12}$ – коэффициент корреляции признаков X_1 и X_2, а $\dfrac{\sum N_g \delta y_g \delta x_{ig}}{N \sigma_0 \sigma_i} = r_{0i}$ – признаков Y и X_i (индекс 0 соответствует Y).

Имеем линейную систему:

$$\begin{cases} a_1 \sigma_1 + a_2 \sigma_2 r_{12} = \sigma_0 r_{01} \\ a_1 \sigma_1 r_{12} + a_2 \sigma_2 = \sigma_0 r_{02} \end{cases}$$

[154]

которая решается очень просто:

$$a_1 = \frac{\sigma_0}{\sigma_1}\frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}, \qquad \text{(III,2,1)}$$

$$a_2 = \frac{\sigma_0}{\sigma_2}\frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2} \qquad \text{(III,2,2)}$$

Теперь полная регрессия Y на X_1 и X_2 имеет вид:

$$y - \overline{y} = \frac{\sigma_0}{\sigma_1}\frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}(x - \overline{x_1}) + \frac{\sigma_0}{\sigma_2}\frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2}(x - \overline{x_2})$$

Допустим, что X_2 фиксировано; обозначая новые средние через $\overline{y'}$ и $\overline{x'}$, получим:

$$y - \overline{y'} = \frac{\sigma_0}{\sigma_1}\frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}(x - \overline{x_1'}) = A(x - \overline{x_1'})$$

Аналогично для регрессии X на Y:

$$x_1 - \overline{x_1'} = \frac{\sigma_1}{\sigma_0}\frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}(y - \overline{y'}) = B(y - \overline{y'})$$

Упражнение 85. Вывести уравнение регрессии X на Y.

Так как произведение коэффициентов регрессии равно квадрату коэффициента корреляции (§ 1, глава III), то, обозначая коэффициент корреляции Y и X_2 при фиксировании X_2 через $r_{01\cdot 2}$, получим: $r^2_{01\cdot 2}=AB$

Отсюда с учетом очевидных обозначений A и B имеем:

$$r_{01\cdot 2} = \frac{r_{01} - r_{02}\cdot r_{12}}{\sqrt{\left(1 - r_{02}^2\right)\left(1 - r_{12}^2\right)}} \qquad \text{(III,2,3)}$$

Обратим внимание на то, что если связи между X_2 и X_1, с одной стороны, и X_2 и Y, с другой, нет, то $r_{01\cdot 2}=r_{01}$, как и следовало ожидать. Полученное выражение является, таким образом, обобщением коэффициента корреляции между двумя признаками (X_1, на Y), если на них влияет третий (X_2). Соотношение (II1,2,3) позволяет определить корреляцию между признаками Y и X_1 при устранении влияния X_2.

Аналогично:

$$r_{02\cdot 1} = \frac{r_{02} - r_{01}\cdot r_{12}}{\sqrt{\left(1 - r_{01}^2\right)\left(1 - r_{12}^2\right)}} \qquad \text{(III,2,4)}$$

$$r_{12\cdot 0} = \frac{r_{12} - r_{01}\cdot r_{02}}{\sqrt{\left(1 - r_{01}^2\right)\left(1 - r_{02}^2\right)}} \qquad \text{(III,2,5)}$$

[155]

Рассмотренные коэффициенты называются коэффициентами корреляции первого порядка (устраняется один признак).

В случае четырех признаков: Y, X_1, X_2, X_3 наряду с коэффициентами рассмотренных типов (r_{01}, r_{12}, $r_{01\cdot 2}$ и т.д.) появляются коэффициенты корреляции второго порядка: например, $r_{01\cdot 23}$ – коэффициент частной корреляции признаков Y и X_1 при устранении влияния X_2 и X_3.

Устраним сперва влияние X_3, вычислив $r_{01\cdot 3}$, $r_{02\cdot 3}$, $r_{12\cdot 3}$, а затем влияние X_2, по ранее рассмотренной схеме. Тогда получим

$$r_{01\cdot 23} = \frac{r_{01\cdot 3} - r_{02\cdot 3}\cdot r_{12\cdot 3}}{\sqrt{\left(1-r_{02\cdot 3}^2\right)\left(1-r_{12\cdot 3}^2\right)}} \qquad \text{(III,2,6)}$$

Можно было сперва устранить влияние X_2, а затем X_3.

Упражнение 86. 1. Записать $r_{01\cdot 23}$ в этом случае. 2. Записать $r_{02\cdot 13}$ и $r_{12\cdot 03}$.

В случае пяти признаков порядок рассмотрения сохраняется, число коэффициентов резко увеличивается. В заключение напомним еще раз, что речь идет о количественных признаках, и все рассмотрение проводилось в предположении линейности связей, а это существенно сужает область применимости данных коэффициентов.

Техника частных корреляций оказывается неприменимой для коэффициентов взаимной сопряженности, ранговой корреляции Спирмена. Однако установлено, что имеет смысл расчет частных корреляций для коэффициента Кендэла. Любопытно, что формулы элиминирования оказываются аналогичными полученным для r. Так, чтобы исключить влияние X_2 на взаимодействие X_1 с Y (случай трех признаков), достаточно рассчитать

$$\tau_{01\cdot 2} = \frac{\tau_{01} - \tau_{02}\cdot \tau_{12}}{\sqrt{\left(1-\tau_{02}^2\right)\left(1-\tau_{12}^2\right)}} \qquad \text{(III,2,7)}$$

и т.д.

3. Множественная регрессия. Случай трех признаков

Частные коэффициенты корреляции, рассмотренные в предыдущем параграфе, выражают связь между результативным признаком («зависимая» переменная) и одним из

[156]

факторов («независимая» переменная) в случае, когда остальные факторы остаются неизменными.

Представляет интерес выявление влияния нескольких признаков (факторов) на результативный. В общем случае это очень сложная задача, которая имеет относительно простое решение, если зависимости линейные.

Рассмотрим для простоты случай трех признаков, который, однако, позволяет понять принцип анализа множественной регрессии в общем случае. Как и в предыдущем параграфе, станем рассматривать признаки Y – результирующий – и факторные X1 и X2. Будем исследовать корреляцию между Y и $U=a_1X_1+a_2X_2$, т.е. признаком, который представляет собой линейную комбинацию факторных. Для этого введем

$$\delta u_g = u_g - \bar{u}, \qquad \text{(III,3,1)}$$

где $g=\overline{1,m}$, как и ранее,

$$u_g = a_1 x_{1g} + a_2 x_{2g}, \qquad \text{(III,3,2)}$$

$$\bar{u} = a_1 \bar{x}_{1g} + a_2 \bar{x}_2 \qquad \text{(III,3,3)}$$

По определению дисперсии

$$\sigma_u^2 = \frac{1}{N}\sum_g N_g(\delta u_g)^2 \qquad \text{(III,3,4)}$$

Естественно определить коэффициент корреляции между *Y* и *U*:

$$R = \frac{\sum_g N_g \delta y_g \delta u_g}{N \sigma_0 \sigma_u} \qquad \text{(III,3,5)}$$

Найдем связь между *R* и r_{ih} *(i, h* =1, 2*)*. Из (III,3,2) и (III,3,3)

$$\delta u_g = a_1 \delta x_{1g} + a_2 \delta x_{2g} \qquad \text{(III,3,6)}$$

Теперь числитель *R* равен

$$a_1 \sum_g N_g \delta x_{1g} \delta y_g =$$

$$= \frac{N \sigma_0^2}{1 - r_{12}^2}\left(r_{01}^2 + r_{02}^2 - 2 r_{01} r_{02} r_{12}\right),$$

если использовать (Ш,2,1), (111,3,2).

[157]

Далее. С учетом (III,3,4) и (III,3,6), а затем (III,2,1) и (III,2,2):

$$\sigma_u = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 - 2a_1a_2\sigma_1\sigma_2 r_{12} = \frac{\sigma_0^2}{1-r_{12}^2}(r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12})$$

Теперь

$$R = \sqrt{\frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1-r_{12}^2}}$$

Вспоминая, что

$$r_{02\cdot 1} = \frac{r_{02} - r_{01}r_{12}}{\sqrt{(1-r_{01}^2)(1-r_{12}^2)}}$$

Мы можем переписать R в виде

$$R = R_{01\cdot 2} = \sqrt{1-(1-r_{01}^2)(1-r_{02\cdot 1}^2)} \qquad \text{(III,3,7)}$$

Индекс у R означает, что коэффициент описывает суммарное влияние признаков X_1 и X_2 на Y.

В случае четырех признаков

$$R_{0\cdot 123} = \sqrt{1-(1-r_{01}^2)(1-r_{02\cdot 1}^2)(1-r_{03\cdot 12}^2)}$$

Заметим, что возможности применения R крайне ограничены, так как линейность встречается в социологии очень редко.

[158]

Г л а в а IV

КЛАССИФИКАЦИЯ СТАТИСТИЧЕСКИХ МЕР ПО УРОВНЮ СОЦИОЛОГИЧЕСКОГО ИЗМЕРЕНИЯ

При изучении статистических мер в предыдущих главах мы рассматривали типы шкал, для которых предназначена мера и условия ее применимости. Сведем теперь эту информацию в единую классификационную схему. Предлагаемая классификация статистических мер по уровням измерения предназначена для того, чтобы облегчить социологу выбор меры, соответствующей полученному им эмпирическому материалу.

Определив тип использованных шкал, исследователь находит соответствующую клетку в приведенных классификационных схемах и определяет меры, допустимые для данных типов шкал. В клетке классификационной таблицы приведено несколько мер для одной и той же ситуации. Например, при изучении связи между двумя номинальными признаками используются коэффициенты *C, T,* T_c и др. Иногда мы даем рекомендации по использованию тех или иных статистических мер (либо в этой главе, либо в главе, где вводилась эта мера), но в некоторых случаях такие рекомендации не даны. Это объясняется недостаточной изученностью вопроса о соотношении между различными статистическими мерами, об их достоинствах, недостатках и типичных ситуациях, в которых их применение наиболее целесообразно[1]. Сам факт использования нескольких мер для одной и той же ситуации говорит о несовершенстве теоретических оснований, используемых для выбора мер.

[159]

[1] Перспективным направлением в изучении мер является сравнительное исследование их поведения путем моделирования таблиц и проведения экспериментов на ЭВМ *(Елисеева И. И., Рукавишников В. О.* Группировка, корреляция, распознавание образов. Ai., 1977, с. 102— 117), но эти исследования не дали пока еще надежных рекомендаций по выбору мер, соответствующих изучаемой ситуации.

По-видимому, имеет смысл рассчитывать все подходящие меры: совпадение результатов, полученных с помощью различных мер, свидетельствует о надежности сделанных выводов. При современной технологии организации обработки социологической информации (гл. VII) это не может сколько-нибудь существенно увеличить временные затраты на обработку. Вообще следует отметить, что встречающиеся в литературе рекомендации по выбору мер статистического анализа, основанные на соображениях удобства расчета, во многом утрачивают свою роль: использование вычислительной техники практически нивелирует различия в сложности расчетов для подавляющего большинства рассмотренных нами показателей.

Из других оснований для выбора статистических мер отметим степень их распространенности: использование распространенных мер повышает возможность сопоставления.

Рассмотрим вкратце основные типы шкал и соответствующие им меры.

Номинальные шкалы

Этот тип шкал представляет собой самый слабый уровень измерения. Такую шкалу называют «примитивной формой» (С. С. Стивенс), «псевдошкалой, которая сама по себе ничего не измеряет» (В. А. Ядов). Тем не менее даже эти самые слабые шкалы позволяют применить довольно значительное число математических процедур для обработки эмпирических данных.

Если рассматривать принадлежность к классу как некоторое свойство, то классы можно интерпретировать как варианты признака. Класс с наибольшей частотой называется модальным. Для номинальных шкал сохраняет смысл понятие «процент».

Мода — единственный вид средней, применимый для номинальных шкал: для этих шкал теряет смысл понятие медианы, ибо медиана — свойство упорядоченного ряда; так как в отсутствии упорядоченности теряет смысл отклонение, нельзя использовать среднее арифметическое, дисперсию. И если роль среднего может играть мода, то в качестве своеобразной меры дисперсии можно использовать нормированную энтропию E (§ 7, гл. II). Другой мерой вариации может служить величина α_k, рассмотренная в § 4 гл. I.

Для изучения связей между признаками, измеренными с помощью номинальных шкал, используется критерий Пир-

сона χ^2, базирующиеся на нем коэффициенты *C, T,* T_c и δ-мера (§ 1, 2 гл. II). Коэффициенты *C* и *T,* как указывалось ранее, для некоторых типов таблиц не достигают единицы. Этот недостаток устранен в коэффициенте Крамера T_c. Наиболее распространенным из них у нас в стране является, пожалуй, коэффициент Чупрова *T.* Для изучения направленных связей используется коэффициент Гудмана *g,* энтропийная мера связи λ и модульный Δ-коэффициент (§ 8 гл. II). Для таблиц 2 х 2 используются коэффициенты ассоциации и контингенции. Если один из признаков измерен с помощью дихотомической номинальной, а второй с помощью порядковой или метрической шкалы, то используются бисериальные коэффициенты корреляции. Для изучения связей между признаками, измеренными с помощью шкал разных уровней, используются меры для более низкого уровня.

Порядковые шкалы

Так как числа, приписанные пунктам порядковой шкалы, отражают отношения равенства (неравенства) попадающих в эти пункты объектов, то для этих шкал, очевидно, применимы все меры, допустимые для номинальных шкал. Сверх того, числа отражают теперь отношения порядка, следовательно, появляются и новые меры. Среди них медиана - позиция, находящаяся в середине ранжированного ряда. При монотонных преобразованиях медианный объект не меняет своего «среднего» положения, хотя и меняется число, описывающее эту позицию. Медиана выступает здесь в качестве показателя средней (центральной) тенденции.

В случае шкал порядка мы, фактически, знаем лишь ранги (последовательность), которые определяют относительную интенсивность качества, но не абсолютную величину его. Теперь не имеет смысла сравнивать интервалы. Поясним это примером. Пусть для объектов *A, B* и *C* имеет место: *X(A)* = 1, *X(B)* = 3, *X*(C) = 7. Ясно, что при этом $X(C) - X(B) > X(B) - X(A)$. После монотонно возрастающего преобразования $X \to X' = \varphi(X)$ такого, что *X′(A)* = 2, *X′(B)* = 10, *X′(C)* = 14, имеем $X'(C) - X'(B) < X'(B) - X'(A)$, т.е. обратное соотношение. Таким образом, строго говоря, статистика, основанная на использовании отклонений (*M*, σ, *D),* не должна

[161]

использоваться при обработке порядковых шкал[2]. Аналогом *M* теперь являются мода и медиана, а аналогом *D* – энтропия и квантили. Пожалуй, следует лишь указать, что в случае, когда при вычислении квантилей приходится прибегать к интерполяции, мы «несколько выходим за пределы экспериментальных свойств шкалы»[3], так как трактуем разность соседних позиций как расстояние. Однако это обстоятельство не приводит к существенным ошибкам и использование квантильной меры более законно, чем, скажем, обычной дисперсии.

Мы уже отмечали, что вычисляя коэффициент ранговой корреляции Спирмена ρ, исследователь использует информацию, которой не располагает (равенству ранговых интервалов, вообще говоря, не отвечает равенство интервалов значений признака). Коэффициент ρ не является мерой, которую – при строгом подходе – можно применять для порядковых шкал. Коэффициент Кендэла, базирующийся на отношениях типа «больше - меньше», выполнимость которых обеспечена эмпирически самой процедурой построения порядковой шкалы, является обоснованной мерой связи между признаками.

Для порядковых шкал можно применять также γ-коэффициент Гудмана, *d*-коэффициент Сомерса и некоторые другие статистические меры.

Интервальные шкалы

Для этих шкал применимы все меры, допустимые для номинальных и для порядковых шкал, которые были рассмотрены выше. Но появляются, конечно, и новые. В частности, в качестве среднего можно вычислить *M,* а для описания вариации - σ.

Рассмотрим объекты A, B и C, которым сопоставлены некоторые числа $X(A)$, $X(B)$ и $X(C)$ в шкале интервалов. Очевидно, имеет смысл, например, утверждать, что $X(A) > X'(B)$, так как и после допустимого преобразования $X \to X' = aX + b (a > 0)$ мы имеем $X'(A) > X'(B)$, что

[162]

[2] Нарушение подобных правил встречается нередко. И не только в социологии. Так, вычисление «средних» баллов успеваемости класса, школы и т.д., фигурирующие в отчетах рай-, гор- и облоно, уязвимо и с точки зрения измерения (выход за пределы свойств шкалы порядка)

[3] *Решлен М.* Измерение в психологии.— В сб.: Экспериментальная психология. М., 1966, с. 211.

вытекает из известных свойств неравенств. Однако утверждения типа *X(A) + X(B) > X*(C) уже оказываются, вообще говоря, лишенными смысла. Действительно, после допустимого преобразования имеем неравенство *X (A) + X (B) + b/a > X(C),* которое истинно для одних *a b* и ложно для других.

Любопытно, что средние значения сравнивать можно. Действительно, рассмотрим две группы: 1) $A_i(i=\overline{1,N})$ и 2) $B_j(j=\overline{1,L})$.

Неравенство

$$\frac{1}{N}\sum_{i=1}^{N}X(Ai)>\frac{1}{L}\sum_{j=1}^{L}X(Bi) \qquad \text{(IV, 1,1)}$$

выполняется и после допустимых преобразований. В самом деле, соотношение

$$\frac{1}{N}\sum_{i=1}^{N}[aX(Ai)+b]>\frac{1}{L}\sum_{j=1}^{L}[aX(Bi)+b]$$

или

$$\frac{a}{N}\sum_{i=1}^{N}X(Ai)+b\frac{N}{N}>\frac{a}{L}\sum_{j=1}^{L}X(Bi)+b\frac{L}{L}$$

выполняется тогда и только тогда, когда выполняется (IV, 11). В данном случае имеет смысл сравнение отношений разностей значений чисел, приписанных объектам: очевидно, что

$$\frac{X(A)-X(B)}{X(C)-X(D)}=\frac{X'(A)-X'(B)}{X'(C)-X'(D)},$$

т.е. сохраняются отношения разностей, или интервалов. Это свойство и дает название шкале.

Заметим (пока формально), что если *b*=0, то при прежнем условии a > 0 имеет смысл не только сравнение интервалов или средних, но и сравнение типа *X(A) + X(B) > X(C),* так как при преобразовании $X \to X' = aX$ сохраняется указанное неравенство. При этом свойством сохранения обладает и отношение:

$$\frac{X(A)}{X(B)}=\frac{X(C)}{X(D)},$$

так как оно (отношение) не изменяется при допустимом преобразовании

$X \to X' = aX\,(a>0)$

[163]

Для изучения связей чаще всего используется коэффициент парной корреляции *r*. Иногда вместо *r* рассчитывается коэффициент ранговой корреляции ρ более удобный тем, что он представляет собой статистику, свободную от распределения, т.е. оперирование с этой величиной не требует каких-либо предположений о форме распределения *X* и *У*[4]. Кроме мер для номинальных и порядковых шкал, используются также коэффициенты частной и множественной корреляции и корреляционное отношение η. Как отмечалось, последний коэффициент обладает тем преимуществом, что позволяет оценивать тесноту не только прямолинейных, но и криволинейных связей.

Шкалы отношений

Это уже рассмотренный нами случай *b*=0. Для таких шкал допустимо вычисление всевозможных статистических мер. Отметим среди новых коэффициент вариации, среднее геометрическое и т.д. Разумеется, эти меры применимы для любых количественных признаков, используемых в социологии (возраст, стаж, заработная плата и т.д.), но не могут быть использованы для качественных признаков.

Так как с повышением уровня шкалы круг допустимых статистических мер расширяется, при переходе к шкале более высокого уровня обычно особое внимание сосредоточивают на «новых» мерах. Однако следует подчеркнуть, что в конкретной ситуации «старые», т.е. применимые и на более низком уровне измерения, меры могут быть более эффективными, чем «новые». Вспомним пример 3 (§ 3 гл. I). Рассматриваемый признак – доход – является количественным, следовательно, возможно вычисление метрического среднего *M,* которое оказывается достаточно большим, но фиктивным, так как изучаемая совокупность была слишком разнородной. Использование мер более низкого уровня – *Me* и *Mo* - позволило лучше понять ситуацию. Это нужно иметь в виду, выбирая статистические меры в каждом конкретном случае.

Приведем сводную таблицу. Ее подлежащим является тип шкалы, сказуемыми – последовательно-базовые эмпирические процедуры построения шкалы, допустимые пре-

[164]

[4] *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи. М., 1973, с. 637.

Таблица 35.

Классификация статистических мер по уровню измерения

Тип шкалы	**Базовая эмпирическаяпроцедура**	**Допустимые преобразования чисел**	Статистические меры		
			центральной тенденции	**вариации**	**связи**
Номинальная	Установление отношения равенстваобъектов	$X \to X' = f(\mathrm{X})$, где $f(X)$ – закон взаимно-однозначногосоответствия	*%Mo*	$\varepsilon\alpha_k$	Q CTT_c $\lambda\delta$ Δg
	↓		↓	↓	↓
Порядковая	Установление отношения последовательности объектов	$X \to X' = \varphi(X)$, где φ(X) — монотонно-возрастающая функция	квантили *Me*	квантильные отклонения	τ парный и частные τ_b, τ_c и d_γ
	↓		↓	↓	↓
Интервальная	Установление равенства интерваловмежду парами объектов	$X \to X' = aX + b\ (a>0)$	*M*	σ C_v	r (парный и частные) ρRη
	↓		↓	↓	↓
Отношений	Установление отношения равенстваотношений паробъектов	$X \to X' = aX\ (\mathrm{a}>0)$	G		
			и любые другие статистические меры		

[165]

образования чисел, не меняющие их (чисел) свойств; статистические меры (средние, вариации, коэффициенты связи).

Примечания:

1. Для построения шкалы данного типа, кроме операции, указанной в соответствующей строке, должны быть эмпирически реализованы операции всех предшествующих типов шкал.

2. Группы преобразований чисел для шкал данного типа входят в группы преобразований шкал всех предшествующих типов, но не наоборот.

3. Для шкал данного типа можно обоснованно применять статистические меры шкал всех предшествующих типов, но нельзя применять меры шкал последующих типов.

Данные положения отражены на схеме с помощью соответствующих стрелок.

Приведем теперь таблицу коэффициентов связи для признаков, измеренных с помощью шкал различных уровней (учитывая, что для интервальных шкал и шкал отношений используются одни и те же меры).

Таблица 36.

Уровни измерения и меры связи между признаками

Тип шкалы *X*	Тип шкалы V		
	Номинальная	Порядковая	Метрическая (интервальная и шкала отношений)
Номинальная	$\Phi\ Q\ C\ T\ T_c$ — $g_{yx}\ g_{xy}\ \delta\ \Delta$ $\lambda_{yx}\ \lambda_{xy}$ ↓	→ — $r_{\rho b}$	→ $r_{rb}\ \eta_{yx}$
Порядковая	— ↓	→ — $\tau\ \gamma$ $d_{yx}\ d_{xy}$ ↓	→ η_{yx} ↓
Метрическая	η_{xy}	η_{xy}	$r\ \rho\ R$

В клетках таблицы, представляющих пересечение строк (уровень измерения признака *X)* и столбцов (признака *Y),* приводятся соответствующие статистические меры связи. Так, на пересечении второй строки (признак *X* — порядковый) и третьего столбца (признак *Y* - метрический) указаны меры связи номинальных и порядковых (с помощью стрелок), порядковых и порядковых шкал, а также η_{yx}, ибо *Y* — метричен. Заметим, что использовать η_{xy} в данной ситуации нельзя, так как неметричен признак *X.*

[166]

Глава V

СТАТИСТИЧЕСКИЕ ВЫВОДЫ: ОЦЕНИВАНИЕ И ПРОВЕРКА ГИПОТЕЗ

1. Генеральная и выборочная совокупность. Оценка ошибки выборки

Объектом социологических исследований обычно являются различные социальные общности. Если изучаются все индивиды данной совокупности, то говорят о сплошном исследовании, если же только часть, то о выборочном. Как правило, социологические исследования носят выборочный характер.

Это связано прежде всего с тем, что экономические и временные ограничения не позволяют провести сплошное исследование (затраты на проведение Всесоюзной переписи, например, составляют десятки миллионов рублей и требуют более 700 тысяч интервьюеров[1])

Но даже в тех случаях, когда сплошное исследование практически осуществимо, зачастую рентабельней проводить выборочное. Его экономичность позволяет увеличить затраты на совершенствование инструмента исследования и компенсировать тем самым падение надежности за счет того, что исследование не сплошное, а выборочное — в итоге исследователь имеет возможность получить более полную и надежную информацию.

Основания, которые позволяют нам по изучению части судить о целом, связаны_с некоторыми вероятностными законами[2]. Если, например, вытаскивать из урны[3], в которой находятся хорошо перемешанные белые и черные камешки (50 белых и 50 черных), 20 камешков, то вероятность того, что нам попадутся все черные камешки очень мала

[167]

[1] Всесоюзная перепись населения — всенародное дело. М., 1978, с. 41.

[2] Примеры с урной широко используются в теории вероятностей и восходят, видимо, к принятой в Древней Греции процедуре голосования.

[3] Примеры с урной широко используются в теории вероятностей и восходят, видимо, к принятой в Древней Греции процедуре голосования.

вероятность того, что первый камешек будет черным – 50/100, что второй — 49/99, так как в урне осталось всего 99 камешков, из них 49 черных; тогда вероятность, что первые два — черные, равна 50/100 * 49/99. Продолжая эти рассуждения, получаем, что вероятность того, что все 20 камешков черные, равна 50/100 * 49/99 * .... * 31/81= 0,00000009. Вероятность того, что одних камешков будет намного больше, чем других, тоже мала. Наиболее часто будет встречаться такая ситуация, при которой число черных камешков приблизительно равно числу белых. Аналогично, если в городе половина населения имеет одни ценностные ориентации, а половина - другие, то маловероятно в выборочном исследовании (при случайном отборе[4]) получить, что лиц с одними ценностными ориентациями намного больше, чем с другими.

Однако осуществить случайный отбор очень трудно. Даже в случае с камешками требуется обеспечить хорошее перемешивание, одинаковые размеры камешков и гарантию, что тот, кто вытаскивает, не видит цвета камешка (примером идеально организованного случайного отбора являются тиражи спортлото). Эксперименты, проведенные с отбором камешков одного цвета, лежащих на столе, показали, что испытуемые непроизвольно выбирают более крупные камешки - они, видимо, чаще попадаются под руку[5]. Несравнимо трудней обеспечить случайный отбор в социологических исследованиях. Широко известен пример неудачного прогноза результатов выборов президента в США в 1936 г.: журнал «Литэрари Дайджест» по телефонным книгам отобрал свыше двух миллионов адресатов, получив тем самым, казалось бы, случайную выборку. По адресам были разосланы открытки с просьбой ответить — Рузвельту или Ландону отдаст свой голос респондент. По результатам опроса журнал предсказал победу с большим перевесом Ландона. Интересно, что социологи Дж. Гэлап и Эл. Роупер правильно предсказали победу Рузвельта, основываясь на анализе в 500 раз меньшего массива – четырех тысяч анкет. Ошибка в прогнозе «Литэрари Дайджест» объясняется тем, что выборка по телефонным книгам не была случайной, она не обеспечивала

[168]

[4] Случайным отбором называют такой, при котором все элементы исследуемой совокупности имеют равную вероятность попасть в выборку.

[5] *Пейте Фрэнк.* Выборочный метод в переписях и обследованиях. М., 1965, с. 34.

равную вероятность попасть в выборку для всех лиц, имеющих избирательное право, так как в 1936 г. телефоны были преимущественно у обеспеченных слоев населения, предпочитавших Ландона.

Свойство выборки отражать характеристики изучаемой совокупности называется *репрезентативностью*. Иногда вместо выборки говорят *выборочная* совокупность, а изучаемую совокупность называют *генеральной*. Можно сказать, что генеральная совокупность – это та, на которую исследователь намерен распространять выводы, сделанные при изучении выборки. Различие характеристик выборочной и генеральной совокупности называют ошибкой репрезентативности. Можно выделить два вида таких ошибок — систематические и случайные.

Систематические ошибки — это ошибки такого типа, как допущенные журналом «Литэрари Дайджест», т.е. некоторое постоянное смещение, которое не уменьшается при увеличении числа опрошенных (если бы журнал опросил не два, а четыре миллиона обладателей телефонов, это не спасло бы его от ошибки).

Случайные ошибки — это те, которые при повторных измерениях изменяются по вероятностным законам. В частности, если мы определяем некоторую характеристику выборки, например, среднее арифметическое, то извлекая все новые и новые выборки того же размера будем получать, что эта характеристика отклоняется то в одну, то в другую сторону от истинного значения (т.е. от значения в генсовокупности) приблизительно с одинаковой частотой и при увеличении числа выборок средняя арифметическая ошибка приближается к нулю. Систематическую ошибку можно устранить, изменяя процедуру формирования выборки; случайная ошибка будет присутствовать всегда, при любом выборочном опросе. Тем не менее систематическая ошибка значительно опасней, так как по выборке ее невозможно оценить. Случайная же ошибка подчиняется определенным законам и поддается оценке. Вообще репрезентативность выборки характеризуется двумя взаимосвязанными параметрами – уровнем ошибки и вероятностью. Говорить о какой-либо выборке, что она репрезентативна, не совсем точно, так как любая выборка имеет определенный уровень репрезентативности (хотя этот уровень может нас совершенно не устраивать). Более точно говорят, что ошибка репрезентативности данной выборки с вероятностью P не превышает Δ (вероятность P называют доверительной).

[169]

Для случайной выборки существуют методы, позволяющие оценить эту ошибку (мы рассмотрим их при изложении способов проверки гипотез). При планировании социологического исследования обычно решают иную задачу — задаются некоторым устраивающим исследователя уровнем точности результата, т.е. допустимой ошибкой и доверительной вероятностью, и определяют для этих параметров необходимый объем выборки. В частности, объем выборки для определения доли некоторого признака X в генсовокупности определяется формулой[6]:

$$n = \frac{1}{\frac{\Delta^2}{t^2\nu(1-\nu)} + \frac{1}{N}},$$

где N - объем генеральной совокупности, n — объем выборки, t - коэффициент, соответствующий доверительной вероятности (см. табл. И Приложения 3; если $n > 60$, то при P=0,954 t=2, а при P=0,997 t=3 и т.д.), ν - доля признака X в генсовокупности, Δ — величина допустимой ошибки (в долях).

Если, например, исследователь хочет получить с вероятностью 0,95 (t = 2) данные о доле признака X в генсовокупности (пусть N = 10000) с ошибкой, не превышающей 5% (Δ = 0,05), и ему известно, что искомая доля составляет приблизительно *20%* (ν = 0,20), то по формуле (V,1,1) получим, что требуется опросить 256 человек *(*n = 256).

Неудобство пользования формулой заключается в том, что она требует хотя бы приближенной информации о доле признака в генеральной совокупности, т.е. как раз о том, что исследователю требуется определить. Чтобы избавиться от этого неудобства, заметим, что при ν=0,5 произведение ν (1 - ν) максимально, следовательно, n тоже максимально. Поэтому если в (V, 1,1) вместо ν подставить 0,5, мы получим формулу, которой можно пользоваться при любых значениях доли признака в генеральной совокупности (объем выборки при этом будет получаться с некоторым запасом). Положив также значение доверительной вероятности равным 0,954, т.е. t = 2, получим

$$n = \frac{1}{\Delta^2 + \frac{1}{N}}, \text{ (V,1,1')}$$

[170]

[6] *Кокрен У.* Методы выборочного исследования. М., 1976, с. 89.

Воспользовавшись этой формулой, определим, как объем выборки зависит от объема генеральной совокупности и от величины допустимой в исследовании ошибки.

Из таблицы видно, что для обеспечения заданной репрезентативности при исследовании города с населением 100 тыс. жителей надо опросить 398 человек, а при исследовании всей страны практически столько же — 400 чел. Для обеспечения одного и того же уровня репрезентативности (5 %) требуется опросить такие доли генсовокупности:

Таблица 37

Зависимость объема выборки от объема генсовокупности при допустимой ошибке 5% доверительная вероятность — 0,954)

Объем генсовокупности	500	1000	2000	3000	4000	5000	10000	100000	Бесконечная
Объем выборки	222	286	333	350	360	370	385	398	400

для N = 500-222 человека, т.е. приблизительно 44% генсовокупности, для N = 5000 - 7,4%, а для N, равном 4 миллионам (например, население Ленинграда) — сотую долю процента. Поэтому изредка встречающиеся в публикациях характеристики выборки типа «было опрошено 15% генсовокупности» или «опрашивался каждый двадцатый школьник» ничего не говорят о репрезентативности выборки. Вообще из таблицы видно, что начиная с некоторого момента увеличение объема генеральной совокупности не оказывает существенного влияния на увеличение объема выборки, поэтому при больших генеральных совокупностях (скажем, при $N > 5000$) величиной $1/N$ в формуле (V,1,1') можно пренебречь. Тогда формула (V,1,1') примет вид:

$n = \frac{1}{\Delta^2}$, откуда $\Delta = \sqrt{\frac{1}{n}}$. Рис. 22, показывающий связь между объемом и ошибкой выборки, может использоваться для принятия решения о требуемом объеме выборки.

При планировании объема выборки следует иметь в виду следующее. Приведенные выше формулы позволяют получить заданную точность при анализе выборки в целом, т.е. если мы не будем расчленять ее на части. Если, например, требуется определить долю лиц, состоящих в браке, для крупного города, то опросив 400 случайным образом

[171]

отобранных человек мы определим искомую долю с ошибкой, не превышающей 5% (с вероятностью 0,954). Но если мы хотим определить эту долю не для всего массива в целом, а для женщин и для мужчин, то нам необходимо, чтобы в выборке было 400 мужчин и 400 женщин, т.е. 800 человек. Чем больше будет дробиться массив при анализе информации, тем больший объем выборки потребуется. Программа

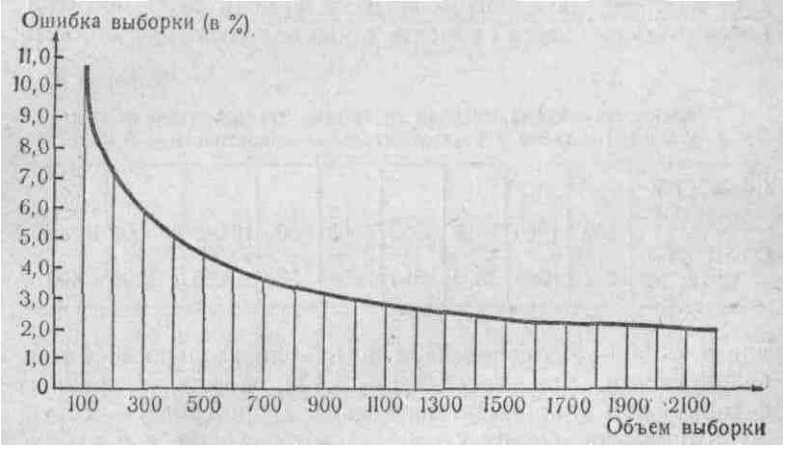


Рис. 22. Зависимость между объемом и ошибкой выборки для $P = 0{,}954$, $p=q=1/2$ и бесконечно большой генеральной совокупности

исследования, план обработки и анализа информации существенно влияют на объем выборки.

Но основная сложность при планировании выборки заключается, пожалуй, в том, что во многих случаях, особенно при крупномасштабных исследованиях, случайную выборку сформировать очень сложно. Так, в проведенном нами репрезентативном исследовании работающего населения г. Киева[7] *(P* = 0,954, Δ не более 3—5%), данные которого уже использовались в примерах, была применена следующая процедура, моделирующая случайный отбор. Из избирательных списков для каждого участка города (всего их более семисот[8]) отбирались с определенным шагом адреса избира-

[172]

[7] Исследование проводилось отделом конкретных социологических исследований Института философии АН УССР (руководитель исследования В. Ф. Черноволенко, проект выборки разработан В. И. Паниотто).

[8] Списки всех идбираюльных участков ранона хранятся в райисполкомах 12-и районов Киева, что очень упрощает риботу.

телей: например, адрес 1-й, 100-й, 200-й и т.д. (если шаг — 100).

При этом могла быть допущена систематическая ошибка такого рода. Фамилии в списках расположены в алфавитном порядке; начиная с 1-го номера, мы почти всегда включаем в выборочный список фамилию на букву А, т.е. в списке доля лиц с фамилией на букву А будет выше, чем в генсовокупности. Если у лиц какой-либо национальности (например, армянской) фамилии чаще начинаются на А, чем у лиц другой национальности, то их доля в выборке выше, чем в генсовокупности.

Поэтому на всякий случай мы «стохастизировали»[9] выбор первой фамилии в списке, сделав его случайным, равномерно распределенным внутри шага выборки (можно, например, начинать с номера равного целой части числа $\frac{k}{7}+1$, где k—номер избирательного участка, изменяющийся от 1 до 700: тогда в 7 избирательных участках фамилии будут отбираться с 1-го номера, в 7 — со 2-го и т.д. до сотого номера.

Поскольку нас интересовало все работающее население, а не только лица, старше 18 лет и занесенные в списки для голосования, опрашивалось не обязательно лицо, указанное в избирательном списке. Список использовался лишь как выборка адресов. При посещении семьи, проживающей по данному адресу, интервьюер переписывал всех проживающих в определенном порядке и по специальной процедуре[10], стохастизирующей выбор, определял, кого надо опросить[11]. Можно показать, что полученная выборка является случайной, для оценки ошибки выборки применима формула (V,1,1).

Но что делать в случае, когда такой список невозможно составить, например, при построении выборки, репрезентативной для Советского Союза? В таких случаях прибегают к многоступенчатому отбору: сначала (первая ступень) из множества всех областей страны отбирают случайным образом области (область в данном случае — это единица).

[173]

[9] От слова стохастический, т.е. случайный.

[10] Использовалась некоторая модификация процедуры Киша. *(Кокрен У.* Методы выборочного исследования. М., 1976, с. 384, 385; *Петренко Е. С., Ярошенко Т. М.* Социально-демографические показатели в социологических исследованиях. М., 1979, с. 96.)

[11] Для реализации случайных отборов иногда используются специальные таблицы случайных чисел (см. табл. М Приложения 3)

отбора на первом шаге). На второй ступени из выбранных областей отбирают районы. На третьей из каждого района — населенные пункты. На четвертой из населенных пунктов — лиц, подлежащих опросу (на первых трех ступенях выбирались единицы отбора, на последней — единицы исследования). При таком построении выборки формула (IV, 1,1) непригодна для оценки ошибки, требуются формулы, позволяющие оценить ошибку, возникающую на каждой ступени[12].

Чтобы оптимизировать процедуру построения выборки, исследователи на первых ступенях отбора используют специальные процедуры. Например, при построении всесоюзной выборки для исследования читателей газеты «Правда»[13] на первой ступени в качестве единиц отбора (или, как их иногда называют, гнезд) использовались области либо республики, если республика не имела областного деления. Всего было выделено 130 гнезд: 71 область РСФСР, 25 областей Украины, 6 областей Белоруссии, 17 областей Казахстана и 11 союзных республик. Прежде чем отбирать гнезда, они были сгруппированы таким образом, чтобы в одну группу попадали территориальные единицы, близкие по урбанизированности, степени развития сельской субкультуры, уровню развития инфраструктуры и промышленного развития области (эти четыре группы переменных описывались 105-ю показателями).

Разбиение на группы (называемое также стратификацией, или районированием) производилось на ЭВМ с помощью так называемого лингвистического метода (это один из методов таксономии, см. главу VI). В результате 130 гнезд были разбиты на 12 страт. Из каждой страты отбиралось некоторое число гнезд (пропорционально численности в данной страте), всего их было отобрано 25. Эта процедура эффективнее случайного отбора 25 из 130 гнезд, так как гарантирует, что в выборку попадут гнезда разных типов, представители всех страт. N и n здесь невелики, поэтому ошибка случайного отбора была бы слишком большой (если бы требовалось отобрать 250 из 1300 гнезд, то можно было бы использовать случайную выборку).

На следующем этапе каждое из гнезд (в данном случае областей) выступало в качестве генеральной совокупности. В каждой области выделялись свои гнезда — города област-

[174]

[12] *Кокрен У.* Методы выборочного исследования. М., 1976, гл. 10.

[13] Территориальная выборка в социологических исследованиях. М., 1980, гл. 3.

ного и республиканского подчинения, районы. Выделенные гнезда группировались в страты, из каждой страты отбиралось определенное число гнезд и т.д. Всего выборка состояла из 6 ступеней (3-я ступень — районы города и села, 4-я — жилищные организации, т.е. ЖЭКи, ЖКК, общежития и т.п.); 5-я — семьи; 6-я — респонденты в семье).

Другим примером многоступенчатой выборки является трехступенчатый отбор респондентов, осуществленный при исследовании городов Татарии[14].

Вообще всякое выборочное исследование может быть охарактеризовано следующими параметрами: числом ступеней, способом выделения гнезд, способом их группировки (стратификации) и способом отбора гнезд на каждой ступени. Очевидно, что во многих случаях способ организации выборки весьма далек от одноступенчатого случайного или даже многоступенчатого случайного отбора. В этих случаях вообще не существует формул для оценки ошибки выборки. Если для некоторых из изучаемых признаков есть контрольные цифры по генсовокупности, полученные из государственной или ведомственной статистики, то можно оценить ошибку выборки по этим признакам. Сопоставляя величину ошибки с той, которая была бы, если бы выборка строилась как одноступенчатая случайная (т.е. с ошибкой, рассчитанной по формуле (V,1,1)), можно оценить отклонения построенной выборки от случайной и внести коррективы в результаты, полученные с помощью формул, основанных на предположении о случайности выборки.

Весь последующий материал этой главы дан в предположении, что из генеральной совокупности извлекается одноступенчатая выборка.

2. Выборочное распределение

Статистический вывод — это некоторое утверждение об изучаемой генеральной совокупности на основании изучения выборки (т.е. рассуждение от частного к общему, индукция). Математическая статистика рассматривает, разумеется, не любые утверждения о генеральной совокупности, а лишь касающиеся числовых характеристик, рассмотренных выше (средние, меры вариации, коэффициенты корреляции,

[175]

и т.п.). Числовые характеристики, описывающие генеральную совокупность, называются *параметрами.* Те же самые характеристики, но рассчитанные для выборки, называются *статистиками.* Мы будем обозначать параметры через M^{Γ}, $(\sigma^{\Gamma})^2$, r^{Γ} и т.д. (генеральное среднее, дисперсия, коэффициент корреляции), а статистики через M^{B}, $(\sigma^{B})^2$, r^{B} и т.д. (выборочное среднее, дисперсия, коэффициент корреляции)[15]. Таким образом, статистический вывод — это утверждение о параметрах генеральной совокупности на основании изучения статистики. Такие утверждения носят вероятностный характер и подразделяются на три вида: статистическое оценивание точечное, статистическое оценивание интервальное и проверка гипотез.

Статистическое оценивание заключается в том, что исследователь по выборке ищет показатель, наиболее близкий к оцениваемому параметру, или интервал, в границах которого с большой вероятностью лежит этот параметр. Другой разновидностью статистического вывода является проверка гипотез: исследователь заранее формулирует некоторое утверждение о параметрах генеральной совокупности (гипотезу), затем оценивает степень соответствия результатов, полученных в выборочном исследовании, сформулированной гипотезе и принимает решение об истинности или ложности гипотезы. Методологию проверки статистических гипотез мы рассмотрим подробнее, что позволит уточнить различие и сходство видов статистического вывода. Сейчас для нас важно, что

---

[14] *Рукавишников В. О., Елисеева И. И.* Проблемы проектирования социологического исследования крупных территориальных объектов.- В кн.: Проектирование и организация выборочного социологического исследования, М., 1977.

[15] В литературе можно встретить также обозначение параметров греческими, а статистик—латинскими буквами. Например, μ, δ^2, π и M^2, s^2, p для среднего, дисперсии и доли признака соответственно.

статистическое оценивание и проверка гипотез основываются на идее так называемого *выборочного распределения* (обращаем внимание читателя на важность этого понятия для понимания сущности статистического вывода).

Рассмотрение выборочного распределения начнем с примера. Предположим, что известно распределение оценок 1000 абитуриентов некоторого вуза на экзамене по математике. Пусть 400 человек получили 2, 200—3, 300—4 и 100—5, полигон распределения приведен на рис. 23. Легко подсчитать, что средний балл равен $M^{\Gamma} = 3{,}1$, а дисперсия $(\sigma^{\Gamma})^2$=0,ll. Насколько вероятно получить в выборке значение, существенно отличающееся от генерального среднего? Определим, например, вероятность того, что для вы-

[176]

борки из 5 человек выборочное среднее M^B будет отличаться от генерального не менее, чем на 0,5, т.е. $|M^Г - M^B| \geq 0{,}5$. С этой целью станем формировать выборки по 5 человек многократно и вычислять для каждой из них средний балл. Тогда отношение таких выборок, для которых $|M^Г - M^B| \geq 0{,}5$, к общему числу извлеченных выборок даст частоту, близкую к искомой вероятности. При увеличении числа выборок эта частота неограниченно приближается к вероятности того, что $|M^Г - M^B| \geq 0{,}5$. Построим полигон рас-

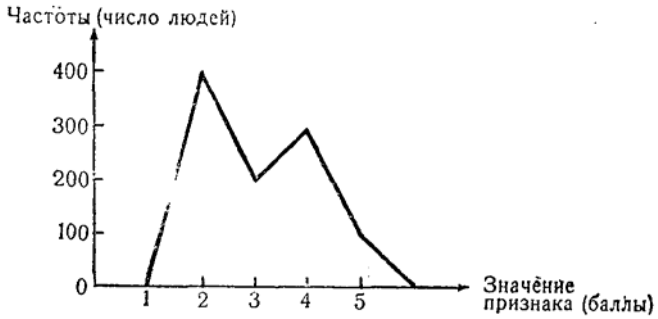


Рис. 23. Пример полигона распределения (оценки по математике)

пределения выборочных средних (по оси абсцисс откладываются значения M^B, по оси ординат — частоты): при увеличении числа выборок до бесконечности получим полигон[16], называемый выборочным распределением статистики M^B (рис. 24). Искомая вероятность равна отношению площади под кривой выборочного распределения, заштрихованный на рис. 24 к площади всей кривой (напоминаем, что площадь под кривой между двумя точками а и & на оси абсцисс равна сумме частот для значений признака, лежащих между *a* и *b*. Зная выборочное распределение, можно решать и другой вопрос — зафиксировать не интервалы изменения статистики, а вероятность, и искать, в каких пределах с заданной вероятностью лежит статистика. Например, можно определить, в каком интервале с вероятностью 0,95 лежит выборочное среднее M^B. Для этого на графике выборочного распределения влево и вправо от генерального среднего $M^Г$ откладывается такой отрезок Δ, чтобы между

[177]

[16] Полигон при этом превратится в плавную кривую.

$M^{Г}$ - Δ и $M^{Г}$ + Δ было заключено 95% площади (на рис. 25 площадь под кривой, лежащей в указанных пределах, заштрихована). Таким образом, зная выборочное распределение, можно определить Δ так, что с вероятностью 0,95 выполняется неравенство: $|M^{Г} — M^{В}|<\Delta$.

Предположим теперь, что мы не знаем генерального среднего $M^{Г}$, но знаем выборочное распределение статистики $M^{В}$

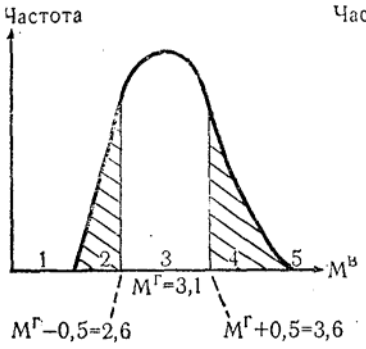


Рис. 24. Выборочное распределение $M^{в}$ для распределения оценок, изображенных на рис. 23

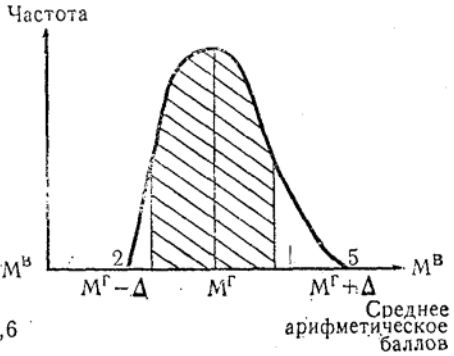


Рис. 25. Выборочное распределение $M^{в}$, доверительный интервал, включающий 95% площади (заштрихована)

и по нему нашли Δ таким образом, что $|M^{Г} — M^{В}|<\Delta$ с вероятностью 0,95. Если мы получим, что в некоторой выборке средний балл абитуриентов равен, например, 3,4, то с вероятностью 0,95 мы можем утверждать, что неизвестное нам генеральное среднее $M^{Г}$ отличается от найденного значения не больше, чем на Δ, т.е. с вероятностью 0,95 $|M^{Г}—3,4|<\Delta$ или 3,4 — Δ < $M^{Г}$ < 3,4 + Δ. Таким образом, мы получили интервальную оценку для неизвестного параметра $M^{Г}$.

Но какой смысл в такой оценке, если для того, чтобы ее получить, надо экспериментально определить выборочное распределение, что практически неосуществимо? Оказывается, однако, что во многих случаях связь между параметром и выборочным распределением статистики носит такой характер, что выборочное распределение статистики можно построить теоретически (более того, для этого часто не требуется никакой информации о форме генерального распределения).

[178]

Рассмотрим этот вопрос более подробно. Пусть из бесконечно большой совокупности с параметрами $M^{Г}$ (среднее арифметическое) и $(\sigma^{Г})^2$ (дисперсия) извлекаются случайные выборки объема *n.* Можно доказать, что при достаточно больших *n* выборочное распределение средних $M^{В}$ будет описываться законом, близким к нормальному, причем среднее арифметическое всех выборочных средних будет равно $M^{Г}$, а дисперсия выборочных средних будет равна $\frac{(\sigma^{Г})^2}{n}$. Это утверждение называется центральной предельной теоремой (слово «центральная» характеризует ее роль в теории статистического вывода). Для предыдущего примера это означает, что выборочное распределение на рис. 25 приближается к нормальному (если бы мы извлекали выборки объема 100, а не 5, то сходство с нормальным распределением было бы значительно большим), причем среднее выборочного распределения равно 3,1, а дисперсия равна 0,11/5 (где 0,11 —дисперсия генерального распределения). Удивительным является тот факт, что распределение выборочных средних близко к нормальному независимо от формы распределения генеральной совокупности. Как уже говорилось (§ 3, гл. I), для нормального распределения существуют таблицы, показывающие, какая доля площади лежит под кривой между любыми двумя точками $M - z\sigma$ и $M + z\sigma$ оси абсцисс (см. табл. А Приложения 3). Известно, например, что в пределах $\sigma^{В}$ от среднего арифметического (т.е. от $M—\sigma^{В}$ до $M+\sigma^{В}$) лежит 68,2% площади кривой; в пределах $2\sigma^{В}$ — 95,4%;

в пределах $3\sigma^{В}$—99,7%. Другими словами, вероятность того, что отобрав из генеральной совокупности *n* единиц и рассчитав $M^{В}$, мы получим, что $|\mathrm{M^{В}} - \mathrm{M^{Г}}| < \frac{2\sigma^{Г}}{\sqrt{n}}$ с вероятностью 0,682; вероятность того, что $|\mathrm{M^{В}} - \mathrm{M^{Г}}| < \frac{\sigma^{Г}}{\sqrt{n}}$ равна 0,954; вероятность того, что $|\mathrm{M^{В}} - \mathrm{M^{Г}}| < \frac{3\sigma^{Г}}{\sqrt{n}}$ равна 0,997. Это и дает возможность по выборке оценивать генеральную совокупность.

Предположим, что в нашем примере мы знаем генеральную дисперсию $(\sigma^{Г})^2 = 0{,}11$, или $\sigma^{Г} = 0{,}33$, но не знаем генерального среднего[17]. Пусть для выборки, состоящей из 5 единиц, оказалось, что $M^{В} = 3{,}3$. Принятая в социологии степень надежности высказываемых утверждений *P* обычно

[179]

[17] Обычно в социологических исследованиях нам неизвестны ни средняя, ни дисперсия генсовокупности. В данном случае мы приняли это допущение, чтобы упростить ситуацию. Ниже будет рассмотрен случаи оценивания, когда неизвестны никакие параметры генсовокупности.

равна[18] 0,954 или 0,997. Для P = 0,954, в силу приведенных выше неравенств: $\left|3{,}3 - M^{\Gamma}\right| < 2 \cdot \frac{0{,}33}{\sqrt{5}} = 0{,}25$, т.е. 3,05 < M^{Γ} < 3,55. Интервал, в который попадает значение оцениваемого параметра, называется доверительным, а вероятность того, что доверительный интервал содержит этот параметр, называется доверительной вероятностью. В нашем примере интервал от 3,05 до 3,55 является доверительным для среднего арифметического генеральной совокупности, построенным с доверительной вероятностью 0,954. Можно также сказать, что (3,05; 3,55) — это 95,4%-и доверительный интервал для M^{Γ} в окрестности точки 3,3.

Может быть, полезной для понимания окажется следующая аналогия. В урне для голосования лежат 20 белых и черных камешков. Предположим, что из теоретических соображений известно, что 19 из них одного цвета, а 1 —другого. Мы вытаскиваем 1 камешек и оказывается, что он белый.Тогда с вероятностью $\frac{19}{20} = 0{,}95$ можно утверждать, что в урне 19 белых и 1 черный камешек. При этом вероятность, что мы ошиблись и в урне 19 черных и 1 белый — тот единственный, который мы вытащили — равна $\frac{1}{20} = 0{,}05$

Упражнение 87. В проведенном нами выборочном исследовании 3500 жителей г. Киева средняя зарплата в выборке равна 150,0 руб. (M^{B} = 150,0). Предположим, что исследования прошлых лет показали устойчивость дисперсии генеральной совокупности и мы полагаем, что она нам известна (пусть $(\sigma^{\Gamma})^2$ == 3700). Найти 95%-й доверительный интервал для средней зарплаты. Ответ: 148 руб.; 152 руб.

Итак, доверительный интервал для неизвестного среднего генеральной совокупности M^{Γ} при известной дисперсии $(\sigma^{\Gamma})^2$ строится следующим образом:

1. Из генеральной совокупности извлекается выборка достаточно большого объема n (100 и более) и рассчитывается среднее $\bar{x}$.

2. Выбирается некоторая доверительная вероятность и по специальной таблице (табл. А Приложения 3) находится коэффициент z, соответствующий этой вероятности.

[180]

[18] Вероятности эти выбраны еще и из тех соображений, что при вероятности 0,954 в неравенстве $|M^{B} - M^{c}| < \sigma^{\Gamma}/\sqrt{n}$, z = 2; при вероятности 0,997: z = 3; часто используются также вероятности 0,95 и 0,99 — при этом z равно 1,96 и 2,58 соответственно.

3. Тогда неизвестный параметр M^{Γ} с вероятностью p лежит в пределах:

$$\text{М}^{\text{в}} - z\frac{\sigma^{\Gamma}}{\sqrt{n}} < M^{\Gamma} < \text{М}^{\text{в}} + z\frac{\sigma^{\Gamma}}{\sqrt{n}}$$

3. Точечное и интервальное оценивание

Предположим, что по выборке нужно найти не интервал, в котором находится параметр, а одно число, которое ближе всего к параметру (его мы хотим использовать для дальнейших вычислений вместо неизвестного параметра). Казалось бы, естественно предположить, что в качестве наилучшего приближения для M^{Γ} следует выбрать M^{B} для $(\sigma^{\Gamma})^2$ величину $(\sigma^{B})^2$ и т.д. Однако это не совсем так. Но прежде, чем рассмотреть этот вопрос подробнее, определим, что мы понимаем под наилучшей оценкой (под оценкой понимается любое число, рассчитанное по выборке и характеризующее параметр).

Принято различать следующие свойства оценок. *Несмещенность* — свойство, состоящее в том, что среднее выборочного распределения оценки равно величине параметра. Это нужно понимать следующим образом: если для оценки некоторого параметра α из генеральной совокупности мы извлечем k выборок объема n, для каждой выборки рассчитаем оценку параметра a_i, и найдем среднее арифметическое этих оценок $\bar{a} = \frac{1}{k}\sum_{i=1}^{k} a_i$, то оно будет близко к параметру α, причем при увеличении k среднее оценок $\bar{a}$ будет стремиться к α. Например, оказывается, что среднее арифметическое выборки $\bar{x} = M^{B}$ является несмещенной оценкой M^{Γ}. Это следует из сформулированной ранее центральной предельной теоремы.

Выборочная дисперсия $(s^{B})^2$ оказывается смещенной оценкой[19] параметра $(\sigma^{\Gamma})^2$, сумма $\frac{1}{k}\sum_{i=1}^{k} (s_i^{B})^2$ при увеличении k стремится к числу, несколько меньшему, чем $(\sigma^{\Gamma})^2$, а именно к $\frac{n-1}{n}(\sigma^{\Gamma})^2$. Если, например, из генеральной совокупности извлекаются выборки объема 2, то оценка $(s^{B})^2$ стремится к величине вдвое меньшей, чем $(\sigma^{\Gamma})^2$, если выборка

[181]

[19] *Кендалл М, Дж., Стьюарт А.* Статистические выводы и связи. М., 1973, с. 18.

объема 3, то $(s^B)^2$ стремится к $\frac{3-1}{3}(\sigma^{\Gamma})^2 = \frac{2}{3}(\sigma^{\Gamma})^2$ и т.д.

При увеличении n это различие существенно уменьшается. Свойство оценки при увеличении объема выборки приближаться к значению оцениваемого параметра называется *состоятельностью* оценки. Таким образом, $(\sigma^B)^2$ является смещенной, но состоятельной оценкой $(\sigma^B)^2$. Отметим, что статистика $\frac{n}{n-1}(\sigma^B)^2$ дает несмещенную оценку для $(\sigma^{\Gamma})^2$.

Обозначим ее через s^2.

Третьим свойством точечных оценок является *эффективность,* мерой которой является дисперсия выборочного распределения статистики. Чем ниже дисперсия, т.е. чем меньше отличаются оценки, полученные в разных выборках, тем выше эффективность. Эти три свойства характеризует качество оценки. Можно показать, что медиана, так же как и среднее арифметическое выборки M^B, является несмещенной оценкой среднего арифметического генсовокупности M^{Γ}. Медиана является также и состоятельной оценкой M^{Γ}. Можно, однако, показать, что дисперсия выборочного распределения медианы приблизительно в полтора раза больше, чем дисперсия выборочного распределения, среднего арифметического, т.е. среднее арифметическое является более эффективной оценкой, чем медиана. Если зафиксировать интервал таким образом, чтобы из 100 выборок в 95-м среднее лежало внутри выделенного интервала, то окажется, что лишь в 61 выборке в этих же пределах лежит и медиана. Оказывается, что выборочное среднее арифметическое обладает наибольшей эффективностью из всех несмещенных и состоятельных оценок M^{Γ} и является, следовательно, наилучшей оценкой для M^{Γ}. Аналогично происходит выбор точечных оценок для других параметров генеральной совокупности.

Вернемся теперь к более важному, как нам представляется, для социологии методу интервального оценивания. Подчеркнем еще раз, что основой интервального оценивания, а также методов проверки гипотез, изложенных в следующих параграфах, является выборочное распределение статистики. Если у читателя нет четкого представления о различии генерального распределения признака и выборочного распределения статистики, то рекомендуем ему внимательно прочесть еще раз страницы, где вводится выборочное распределение.

[182]

Упражнение 88. А. Сформулировать определения: 1) генеральное распределение признака; 2) выборочное распределение признака; 3) выборочное распределение статистики.

Б. Пусть из некоторой бесконечной генеральной совокупности извлечено 1000 выборок по одному элементу каждая, измерено значение признака X для каждого элемента и построено распределение. Какое из названных в п. *А* распределений мы получим?

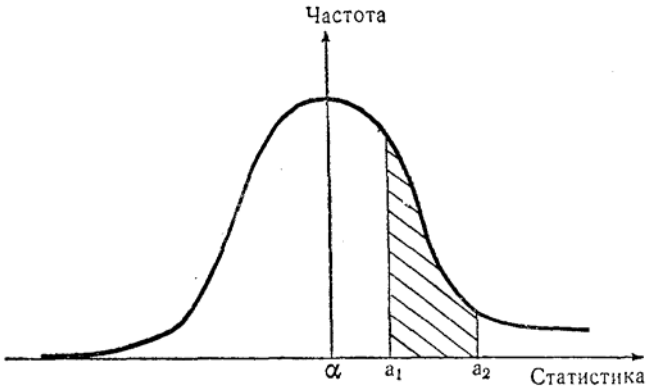


Рис. 26. Распределение статистики a для выборок объема n из генеральной совокупности с параметром α

В. Если в предыдущем случае в каждой из выборок извлекать не по одному, а по два элемента (получим X_1 и X_2) и затем построить распределение. Какое из названных распределений мы получим?

Г. В предыдущем случае в каждой выборке будем вычислять $Y = 1/2\ (X_1 + X_2)$ и построим распределение Y.

Какое из названных распределений мы получим?

Ответ: Б и В — выборочное распределение признака, Г —распределение статистики.

В предыдущем параграфе мы рассмотрели, как строится интервальная оценка генерального среднего. Аналогично получаются интервальные оценки и других параметров. В общем виде схему интервального оценивания можно представить следующим образом. Пусть дана генеральная совокупность с некоторым неизвестным параметром и, значение которого требуется оценить (это может быть доля признака, среднее, мера вариации, коэффициент корреляции, разность средних, разность коэффициентов корреляции и т.п.).

[183]

Для этого выбирается оценка параметра α — некоторая статистика (как правило, наилучшая, т.е. несмещенная, состоятельная и максимально эффективная оценка). Нужно найти выборочное распределение этой статистики. Для этого исследователь не извлекает выборки объема *n* из генеральной совокупности, не вычисляет статистику a_i, *(i* — номер выборки) и не строит распределение значений a_i, а определяет, каким будет распределение, теоретически. Пусть такое распределение найдено (рис. 26). Важно отметить, что статистическая теория дает также возможность определить по выборочному распределению, каково значение неизвестного параметра. Например, как нам уже известно, если построено выборочное распределение среднего арифметического, то среднее выборочного распределения (т.е. среднее средних арифметических каждой выборки) равно параметру. Это свойство выполняется во всех случаях, когда в качестве статистики выбирается несмещенная оценка параметра генеральной совокупности (по определению несмещенности).

Далее полученное распределение табулируется: для каждого значения Δ, взятого с некоторым шагом, определяют долю площади, которая лежит под кривой выборочного распределения[20] от α — Δ до α + Δ. Эта доля, т.е. отношение этой площади ко всей площади, лежащей под кривой, равна вероятности того, что в выборке значение статистики a_i, будет больше α — Δ и меньше α + Δ. Например, если выборочное распределение нормально, то по таблице А Приложения 3 можно определить, что от -0,1 до 0,1 лежит 0,00798% площади кривой, от -1,2 до 1,2 — 0,76986% и т.д. (в таблице приведены значения для случая, когда $\alpha = 0$, $\sigma = 1$; если же $\alpha \neq 0$, $\sigma \neq 1$, то в пределах от $\alpha - 0{,}1\sigma$ до $\alpha + 0{,}1\sigma$, лежит 0,00798% площади кривой, от $\alpha - 0{,}2\sigma$ до $\alpha + 0{,}2\sigma$; лежит 0,76986% площади кривой и т.д.).

По таблице можно также определить долю площади, лежащей между любыми двумя точками a_1 и a_2 (т.е. вероятность того, что значение статистики a_i в выборке будет лежать в интервале от a_1 до a_2). Для этого сначала определяют доли площади, лежащие от 0 до a_1 и от 0 до a_2 (площадь, лежащая между 0 и *a*, это половина площади, лежащей между — *a* и *a)*. Тогда площадь между a_1и a_2, определится

[184]

[20] Иногда табулируется площадь, лежащая между 0 и Δ или между Δ и ∞ - это зависит от выборочного распределения и от конкретной таблицы.

как разность площади, лежащей от 0 до a_2, и площади, лежащей от 0 до a_1.

Упражнение 89. Пусть выборочное распределение переменной x описывается нормальным распределением со средним 0 и дисперсией 1 (т.е. таблицей А Приложения 3). Определите, какова вероятность, что в выборке мы получим значение, лежащее между а) -2 и 2; б) -2,58 и 2,58; в) -1,4 и 1,4; г) 1 и 1,4; д) -1 и 1,4.

Вся описанная в п. 1 работа выполняется не социологом, а статистиком, для большинства стандартных случаев она уже проделана и наиболее часто встречающиеся выборочные распределения протабулированы (это нормальное распределение, распределение χ^2 (хи-квадрат), F — распределение Фишера и t — распределение Стьюдента, см. таблицы А, Б, И и Л Приложения 3). Поэтому социолог должен лишь определить, каким выборочным распределением описывается его статистика, т.е. какую из таблиц он должен выбрать (изложению этого и посвящены последующие параграфы этой главы). Таким образом социологу нужно:

1. Определить, какой статистикой следует пользоваться и найти соответствующую таблицу.

2. Задавшись некоторой доверительной вероятностью (например, 0,95), по выбранной таблице для заданной вероятности определить такое число Δ, чтобы в пределах от α — Δ до α + Δ лежало 95% площади кривой. Это означает, что с вероятностью 0,95 любое выборочное значение a лежит в этих пределах, т.е.

$|\alpha - a| < \Delta$

или: $a - \Delta < \alpha < a + \Delta$ (V,3,1)

3. Из генсовокупности извлекается случайная выборка и вычисляется значение статистики а. Из неравенства (V,3,1) тогда следует, что $(a — \Delta, a + \Delta)$ и есть искомый 95%-и доверительный интервал.

В последующих параграфах при изложении методов проверки гипотез будут рассмотрены конкретные случаи отыскания доверительных интервалов для процентов, средних, коэффициентов корреляции и т.п.

4. Проверка статистических гипотез

Предположим, что исследователь провел на некотором крупном предприятии сплошной опрос и оказалось, что средний балл удовлетворенности трудом равен α (пусть, например, $\alpha = 0{,}43$), а дисперсия равна $(\sigma^Г)^2$ (пусть $(\sigma^Г)^2 = 1{,}26$,

[185]

т.е. $\sigma^{Г}$ = 1,12). По рекомендациям, разработанным социологом, на предприятии были проведены мероприятия, направленные на повышение удовлетворенности работников трудом. Через год для проверки эффективности мероприятий исследователь провел выборочное исследование объема *n* (пусть *n* = 100) и получил, что средний балл удовлетворенности в выборке равен *b,* причем *b* несколько выше, чем α. (например, *b=0*,68), а дисперсия не изменилась ($(\sigma^{B})^2$ =1,26). Возникает вопрос, произошли ли какие-нибудь изменения

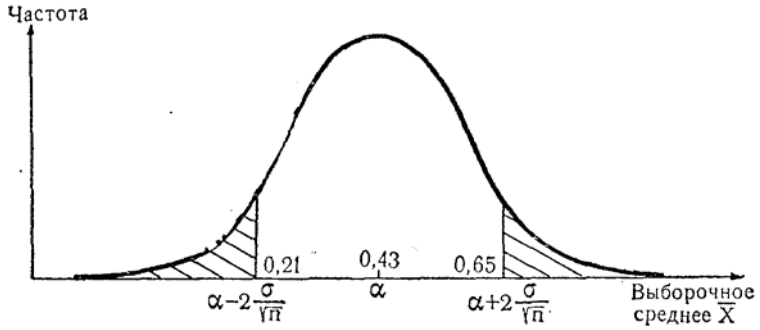


Рис. 27. Выборочное распределение: площадь, соответствующая вероятности совершить ошибку I рода (заштрихована)

на предприятии или тот факт, что $\beta > \alpha$, является случайным[21], и проведя сплошной опрос, мы никаких изменений не обнаружили бы (т.е. новое значение удовлетворенности ($\beta = \alpha$)? Другими словами, можно высказать две гипотезы о неизвестном параметре генеральной совокупности (обозначим их через H_0 и H_1):

1. H_0: $\beta = \alpha$. (так называемая, нулевая гипотеза)
2. H_1: $\beta \neq \alpha$ (альтернативная)

Какая из гипотез более обоснована? Для принятия решения в данном случае поступают следующим образом. Предположим, что гипотеза H_0 верна (т.е. (β = 0,43), а дисперсия $(\sigma^{Г})^2$ равна 1,26, как и ранее. Тогда выборочное распределение описывается нормальной кривой со средним $\beta = \alpha$ (т.е. 0,43) и дисперсией $(\sigma^{B})^2 == \frac{(\sigma^{Г})^2}{n}$ (рис. 27). Как известно, 95,4% площади нормального распределения лежит в пре-

[186]

[21] Введенные обозначения связаны с тем, что β — неизвестный параметр генеральной совокупности, а *b* — статистика.

делах двух среднеквадратических отклонений от среднего, а так как среднее квадратическое отклонение выборочного распределения в данном случае равно $\frac{\sigma^{\Gamma}}{\sqrt{n}}(\frac{1,12}{\sqrt{100}} = 0,11)$, то, следовательно, 95,4% выборочных средних лежит в пределах от $\alpha - \frac{2\sigma}{\sqrt{n}}$ до $\alpha + \frac{2\sigma}{\sqrt{n}}$ (т.е. от 0,43 -0,22 = 0,21 до 0,43 + 0,22 = 0,65) и b = 0,68, лежит вне интервала (0,21; 0,65). Если H_o верна, то вероятность получить значение «b» вне этого интервала равна 0,046 (4,6%). Поэтому в данном случае разумно отклонить гипотезу H_o и принять гипотезу H_1 (риск, что мы совершили ошибку и что гипотеза H_0 верна составит лишь 4,6%). Такая ошибка — отклонить гипотезу H_0, когда она верна — называется *ошибкой I рода.*

Вероятность совершить ошибку I рода называется уровнем значимости (эту вероятность выражают и в процентах). Поэтому эквивалентным вышеприведенному является утверждение «гипотеза H_0 равенстве средней удовлетворенности работников предприятия до и после проведенных мероприятий отклоняется на уровне значимости 0,046». Чаще всего в социологических исследованиях задают 5- или 1%-ный уровень значимости. В нашем случае доверительный интервал для проверки на 1%-ном уровне значимости равен $\alpha - 2,58\frac{\sigma^{\Gamma}}{\sqrt{n}}; \alpha + 2,58\frac{\sigma^{\Gamma}}{\sqrt{n}}$, т.е. (0,14; 0,72). Если полученное значение b было бы меньше 0,14 или больше 0,72, то мы с большей уверенностью могли бы утверждать, что произошли изменения, и отклонить гипотезу H_0 на 1 %-ном уровне.

Принятая при проверках гипотез терминология включает также понятия «критическая область» и «критическая точка». *Критическая область* — это те значения статистики, при которых отвергается гипотеза H_0 (в каком-то смысле это понятие дополнительное к понятию доверительного интервала). Так, для 5%-го уровня значимости критической областью при проверке гипотезы H_0 в приведенном примере являются значения, лежащие вне интервала (0,21; 0,65), т.е. значения, которые меньше 0,21 и больше 0,65 (площадь под кривой выборочного распределения, лежащая в этих пределах, на рис. 27 заштрихована). Точки, отделяющие критическую область от области принятия гипотезы (т.е. 0,21 и 0,65), называются *критическими.*

[187]

В этих терминах процесс проверки гипотез описывается следующим образом. Исследователь формулирует гипотезы H_0 (нулевую) и H_1 (альтернативную) и определяет, каким будет выборочное распределение статистики, служащей для оценки параметра, о котором сформулирована гипотеза H_0 , если предположить, что она верна.

Следующий шаг — выбор уровня значимости и определение критической области для этого выборочного распреде ления

Таблица 38

Ошибки при проверке статистических гипотез

Исследователь принял решение	В действительности	
	H_0 верна (т.е. H_1 неверна)	H_1 верна (т.е. Но неверна)
Отклонить H_0 (т.е. принять H_1)	Ошибка I рода Вероятность (т.е. уровень значимости) q	Правильное решение Вероятность (т.е. мощность) $1 — p$
Отклонить H_0 (т.е. принять H_1)	Правильное решение Вероятность $1 — q$	Ошибка II рода Вероятность p

при данном уровне значимости. И последний шаг — проведение выборочного исследования и определение выборочного значения статистики: если полученное значение попадает в критическую область,— гипотеза *Но* отвергается и принимается альтернативная гипотеза H_1, в противном случае, говорят, что выборочное значение статистики попало в область принятия гипотезы и гипотеза принимается.

Важно отметить, что в принятой схеме проверки нуль-гипотеза и альтернативная неравноправны. Хотя в учебниках по статистике[22] нуль-гипотезу иногда определяют просто как любую проверяемую гипотезу, поменять ее местами с альтернативной гипотезой не всегда возможно. Если бы в приведенном нами примере в качестве нуль-гипотезы была бы принята гипотеза о неравенстве средних ($\beta \neq \alpha$), то мы столкнулись бы с серьезными трудностями.

Действительно, для проверки такой гипотезы надо пред-

[188]

[22] *Гласс Дж, Стэнли Дж.* Статистические методы в педагогике и психологии. М., 1976, с. 254; Статистические методы анализа информации в социологических исследованиях. М., 1979, с. 81.

положить, что она верна и определять затем выборочное распределение статистики b для проверки этого утверждения. Но таких распределений может быть не одно, как в случае, если $(\beta = \alpha$, а бесконечно много (ведь существует бесконечно много β не равных α)! Поэтому определить критическую область для проверки гипотезы H_1 в данном случае очень сложно. Ошибка отклонить альтернативную гипотезу H_1 при условии, что она верна, называется *ошибкой II рода,* а вероятность не допустить эту ошибку называется мощностью. Соотношение введенных понятий хорошо видно из таблицы 38.

Мы не будем рассматривать методы расчета ошибок второго рода из-за их сложности и потому, что они практически не используются социологами (по крайней мере, нам неизвестна ни одна отечественная публикация, в которой рассчитывалась бы ошибка II рода). Однако анализ различий этих видов ошибок позволяет лучше понять методы статистической проверки гипотез. Почему оценка вероятности ошибки I рода получила гораздо более широкое применение в социологических исследованиях (и, пожалуй, в исследовательской работе вообще), чем оценка вероятности ошибки II рода?

Целью исследователя является поиск различных закономерностей, связей изучаемых явлений, различий между изучаемыми группами респондентов и т.п. Используя математическую статистику, он стремится показать, что между изучаемыми переменными есть связь, что рассчитанный коэффициенты корреляции значимо отличаются от нуля, что между двумя процентами есть различия и т.д., т.е. исследователь стремится показать, что обнаруженные им различия в выборочных данных не обусловлены игрой случая, что в действительности (т.е. в генеральной совокупности) они существуют.

Путь, предлагаемый математической статистикой — это доказательство от противного: предположим, что никаких различий нет, т.е. проценты равны, средние арифметические не различаются между собой, коэффициент корреляции равен нулю и т.п. (именно такого рода гипотезы, а не любые произвольные формулируют в качестве нуль-гипотезы).

Если даже в действительности различий нет, то из-за случайных обстоятельств в выборке их можно все же получить — статистика позволяет теоретически оценить, насколько большими могут быть эти случайные различия, например, показать, что с вероятностью 0,99 при отсутствии различий в действительности, различия в выборке за счет

[189]

случайностей не могут превышать некоторого числа *k* (критическая точка).

Если оказывается, что в эмпирическом исследовании получено различие большее *k,* то это противоречит гипотезе об отсутствии различий и, следовательно, гипотеза H_1 принимается с определенной вероятностью (скажем, 0,99). При этом, естественно, есть риск, что исследователь ошибся и в действительности различий нет (вероятность этого 0,01). Это и есть ошибка I рода.

Допустить такую ошибку — все равно, что заявить, об открытии, которое оказывается фикцией. В приведенном выше примере, если бы социолог допустил ошибку I рода, то это значило бы, что он утверждал, что предложенные им мероприятия эффективны, хотя в действительности это не так. Что же касается ошибки II рода, т.е. принятие нулевой гипотезы — хотя в действительности есть различия — то она выглядит значительно менее неприятной.

К каким последствиям она приведет? К тому, что исследователь будет продолжать работу, увеличивать объем выборки, точность инструмента исследования и если различия есть, то он их в конце концов обнаружит. Принятие нулевой гипотезы означает, скорее, не строгое отсутствие различий, а то, что либо различий нет, либо они невелики. Здесь, как и в методе доказательства от противного, если мы получили противоречие сделанному допущению, то допущение неверно, но если мы путем некоторого рассуждения не получили противоречия, то это еще не значит, что допущение верно – другой ход рассуждений может привести к противоречию.

Такая ситуация типична для проверки научных теорий вообще — отсутствие фактов, противоречащих теории, еще не означает, что она верна — научный поиск может в конце концов привести к открытию нового факта, который противоречит теории, что потребует разработки новой теории, включающей и этот факт. Таким образом, принятие нулевой гипотезы означает не подтверждение ее, а неопровержение, поэтому ошибка II рода не так опасна, как ошибка I рода. Если ошибка I рода приводит к получению ложных фактов и вносит шум, помехи в научную информацию, то ошибка II рода несколько отдаляет получение нового факта и увеличивает затраты, но стимулирует исследователя к поиску новых, более совершенных и точных методов.

Вероятности ошибок I и II рода связаны обратной зависимостью; уменьшение вероятности ошибки I рода (умень-

[190]

шение уровня значимости) приводит к увеличению вероятности ошибки II рода. К счастью, с увеличением объема выборки падают вероятности ошибок I и II рода.

Таким образом, не вызывает сомнения важность оценки вероятности ошибок I рода. К сожалению, в социологической литературе часто встречаются работы, авторы которых довольно легкомысленно склонны трактовать даже незначительные различия в полученных ими эмпирических данных, не оценивая вероятности ошибок I рода. Нам представляется обязательным расчет уровней значимости — это дисциплинирует исследователя и позволяет как сквозь сито просеять эмпирические данные, оставить лишь наиболее надежные факты (напомним, что даже 5%-ный уровень значимости в среднем в одном случае из 20 дает ложный факт).

Перейдем теперь к рассмотрению конкретных, наиболее распространенных случаев статистического вывода, в частности, проверки гипотез и построения доверительных интервалов. Изложение ведется в такой последовательности:

1) формулирование нулевой и альтернативной гипотез;

2) определение выборочного распределения для проверки нулевой гипотезы[23]; 3) определение критических значений для проверки гипотез; 4) построение доверительных интервалов; 5) пример; 6) упражнение.

5. Значимость различий долей (процентов)

Эта задача, по-видимому, чаще всего встречается в социологических исследованиях: имеются две генеральные совокупности, из одной извлечена выборка объема n_1, из другой — независимая выборка объема n_2. Оказалось, что доля некоторого признака в одной выборке ν^{β}_1, а в другой ν^{β}_2. Возникает вопрос, не обусловлено ли различие ν^{β}_1 и ν^{β}_2 случайными факторами, т.е. различаются ли доли этого признака в генеральных совокупностях? В реальных исследованиях чаще встречается ситуация, когда извлекается одна выборка, которая затем разбивается на группы (например, по полу, по характеру труда и т.п.), и ставится задача определить, различаются ли выделенные группы по доле изучаемого признака (например, различаются ли

[191]

[23] В зависимости от вида распределения используется то или. иное обозначение для статистики: z — при нормальном распределении, t и F — при распределении Стьюдента и Фишера соответственно.

мужчины и женщины по доле рационализаторов). В этом случае можно считать каждую из групп выборкой из своей генеральной совокупности (например, при городском выборочном опросе мужчины, попавшие в выборку, представляют генеральную совокупность «мужское население города», а женщины — соответственно «женское население города»).

Итак, предположим, что даны две бесконечные генеральные совокупности (будем считать, что генеральная совокупность бесконечна, точнее, что мы имеем право пользоваться формулами, выведенными для бесконечной генеральной совокупности, если объем выборки *n* составляет менее 5% от объема генеральной совокупности *N,* но *N* при этом не менее 1000). Из первой генеральной совокупности извлечена выборка объема n_1, из второй — объема n_2, доля признака *X* в первой выборке ν^B_1, во второй ν^B_2, неизвестные доли признака *X* в генеральной совокупности составляют $\nu^Г_1$ и $\nu^Г_2$ соответственно.

1. Формулируем гипотезы:

1) H_0: $\nu_1^Г = \nu_2^Г$

2) H_1: $\nu_1^Г \neq \nu_2^Г$

2. Пусть $n_1 \geq 50, n_2 \geq 50, n_1\nu_1^B > 5, n_2\nu_2^B > 5, \quad n_1(1-\nu_1^B) > 5, \quad n_2(1-\nu_2^B) > 5.$ Тогда, если верна гипотеза H_0, приведенная ниже функция от $(\nu_1^B - \nu_2^B)$ имеет нормальное распределение с нулевым средним и единичным среднеквадратичным отклонением:

$$z = \frac{\left|\nu_1^B - \nu_2^B\right|}{\sigma_z}, \text{ (V,5,1)}$$

где $\sigma_z = \sqrt{\nu^B(1-\nu^B)\cdot(\frac{1}{n_1}+\frac{1}{n_2})},$

$$\nu^B = \frac{\nu_1^B n_1 + \nu_2^B n_2}{n_1 + n_2}$$

3. Задавшись некоторым уровнем значимости, по таблице А Приложения 3 определяем критические точки. Например, при 5%-м уровне значимости - z = 1,96, т.е. критические точки - 1,96 и +1,96, а на 1%-ном - 2,58 и +2,58. Область, лежащая между критическими точками, является областью принятия гипотезы, а вне этих точек — критической областью. Если, например, подставив полученные в эксперименте данные в формулу (V,5,1), мы по-

лучили, что -1,95< z <1,96, то гипотеза *Но* принимается, а если $z \leq 1{,}96$ или $z \geq 1{,}96$, то гипотеза отвергается на 5%-м уровне значимости.

4. Доверительные интервалы для доли признака в каждой из выборок можно найти даже в случае конечной генеральной совокупности. Если, например, первая из рассмотренных генеральных совокупностей состоит из N_1 единиц, то доверительный интервал задается формулой (V,5,2)

$$\left(\nu_1^B - \frac{1}{2n_1}\right) - z\sqrt{\frac{\nu_1^B(1-\nu_1^B)}{n_1}\cdot\frac{N_1-n_1}{N_1-1}} < \nu_1^\Gamma < \left(\nu_1^B - \frac{1}{2n_1}\right) + z\sqrt{\frac{\nu_1^B(1-\nu_1^B)}{n_1}\cdot\frac{N_1-n_1}{N_1-1}}$$

где z — коэффициент, который определяется по таблице А Приложения 3 (например, при 95%-ной доверительной вероятности, т.е. при 5%-ом уровне значимости, z = 1,96). Аналогично определяются доверительные интервалы для доли признака в любой генеральной совокупности[24].

Что же касается доверительного интервала разности долей, то он определяется формулой

$$\left(\nu_1^b - \nu_2^b\right) - z\sigma_z^b < \left(\nu_1^\Gamma - \nu_2^\Gamma\right) < \left(\nu_1^b - \nu_2^b\right) + z\sigma_z^b \text{ (V,5,3)}$$

5. Рассмотрим *пример № 29.* Исследование общественного мнения об Олимпиаде-80 в Москве[25] показало, что по мере приближения к началу Олимпийских игр интерес москвичей к ним увеличивался: в двух исследованиях, проводившихся с интервалом в полгода, доля ответивших, что вопросы, связанные с подготовкой и проведением Олимпиады, их не интересуют, уменьшилась с 9% до *4°о.* Проверим на 5%-ом уровне значимости, что увеличение интереса к Олимпиаде действительно имело место. В указанной работе данные о выборке приводятся по второму исследованию *(n_2*= 919), о первом сказано лишь, что оно проводилось в двух районах Москвы. Предположим, что первая выборка включала 300 опрошенных *(n_1*= 300), $v_1^B = 0{,}09$, $v_2^B = 0{,}04$. Ясно, что $n_1 = 300 > 50$, $n_2 = 919 > 50$, $n_1 v_1^B = 27 > 5$,

[193]

[24] Приближенно 95-й и 99%-и доверительные интервалы можно определить, не проводя каких-либо расчетов, по таблице 3 Приложения 3.

[25] *Коробейников В. С., Воинова В. Д., Токаровский Г. Д.* Общественное мнение об Олимпиаде в Москве.— Социологические исследования, 1980, .№ 2, с. 153.

$n_2\nu_2^B = 36{,}8 > 5, n_1(1-\nu_1^B) = 300 \cdot 0{,}91 = 273 > 4$ и, наконец, $n_2\left(1-\nu_1^B\right) = 919 \cdot 0{,}96 = 882 > 5$.

Проводим вычисления по формуле (V,5,1):

$$\nu^B = \frac{0{,}09 \cdot 300 + 0{,}04 \cdot 919}{300 + 919} = 0{,}052$$

$$\sigma_z^B = \sqrt{0{,}052 \cdot 0{,}948\left(\frac{1}{300} + \frac{1}{919}\right)} = 0{,}0148$$

$$z = \frac{0{,}09 - 0{,}04}{0{,}0148} = 3{,}38$$

Как видим, полученное значение больше, чем 1,96, поэтому нулевая гипотеза, заключающаяся в том, что за полгода никаких изменений в отношении москвичей к Олимпиаде не произошло, отвергается на 5%-ом уровне. Поскольку полученное значение больше, чем 2,58, гипотеза отвергается и на 1 %-ом уровне, наличие изменений можно считать доказанным.

Найдем теперь доверительные интервалы. Пусть доверительная вероятность 0,95 *(z* = 1,96). Тогда нижняя граница доверительного интервала доли лиц, не проявляющих интереса к Олимпиаде в первом опросе, равна (заметим, что поскольку генеральную совокупность — жители Москвы – можно считать бесконечной, выражение $\frac{N-n}{n-1}$ в формуле (V,5,2) равно1):

$$\left(0{,}09 - \frac{1}{2 \cdot 300}\right) - 1{,}96\sqrt{\frac{0{,}09 \cdot 0{,}91}{300} \cdot 1} = 0{,}056$$

Верхняя граница соответственно равна

$$\left(0{,}09 - \frac{1}{2 \cdot 300}\right) + 1{,}96\sqrt{\frac{0{,}09 \cdot 0{,}91}{300} \cdot 1} = 0{,}124$$

Таким образом, с вероятностью 0,95 доля лиц, не проявляющих интереса к Олимпиаде при первом опросе, лежит в пределах от 0,056 до 0,124 (или от 5,6% до 12,4%). Аналогично рассчитываем, что во втором опросе соответствующая доля лежит в пределах от 2,7% до 5,3%. Доверительные пределы для разности долей ищем по формуле (V,5,3)

(0,09 - 0,04) - 1,96 • 0,0148 < v_1^B — v_2^B < (0,09 — 0,04) + 4-1,96 • 0,0148.

[194]

Таким образом, с вероятностью 0,95 величина, на которую снизилась доля лиц, не проявляющих интерес к Олимпиаде, лежит между 2,1% и 7,9%.

6. *Упражнение 90.* В проведенном нами почтовом опросе работающего населения г. Киева были получены следующие данные (табл. 39).

Таблица 39

Семейное положение мужчин и женщин г. Киева (занятое население)

Пол	Число опрошенных	Семейное положение			Всего
		женат (замужем)	Сейчас не женат (не замужем), но ранее был (а)	Не женат (не замужем) и не был (а)	
Мужчины	1150	85,7%	4,3%	9,9%	100%
Женщины	1365	71,0 %	17,2%	11,8%	100%

Проверить на 5%-ном уровне значимости, отличаются ли доли женатых мужчин и замужних женщин и построить 95%-ные доверительные интервалы для долей женатых мужчин, замужних женщин и для разности этих долей.

6. Значимость различий средних арифметических

Даны две независимые выборки объема n_1 и n_2 из бесконечных генеральных совокупностей с неизвестными дисперсиями и неизвестными средними $M_1^{Г}$, $M_2^{Г}$. По каждой из выборок рассчитаны средние $M_1^{В}$ и $M_2^{В}$ и оценки[26] дисперсий s_1^2 и s_2^2

1. Формулируем гипотезы

$H_0 = M_1^{Г} = M_2^{Г}$

$H_1 = M_1^{Г} \neq M_2^{Г}$

[195]

[26] Несмещенной оценкой генеральной дисперсии является не выборочная дисперсия, а величина $s^2 = \frac{n}{n-1}(\sigma^{В})^2$, где $(\sigma^{В})^2$ – выборочная дисперсия. Поэтому в формулах для проверки гипотез обычно используется s (впрочем, при n>100 различие s и $\sigma^{В}$ несущественно).

2. Рассчитываем показатели

$$t = \frac{\left| M_1^B - M_2^B \right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{(V,6,1)}$$

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 + 1}} \quad \text{(V,6,2)}$$

Если верна гипотеза H_0, то показатель t, рассчитанный по формуле (V,6,1), имеет так называемое t-распределение Стьюдента с f степенями свободы.

3. По таблице И Приложения 3 определяем критические точки: задаемся некоторым уровнем значимости, например 5%-ным, и находим соответствующий столбец; берем ближайшее к полученному по формуле (V,6,2) значению f целое число и находим соответствующую строку — на пересечении найденной строки и столбца стоит критическое значение $t_{кр.}$ Например, для критерия значимости 5% и $f = 2$, $t_{кр} = 4{,}30$. Если полученное по формуле (V,6,1) значение t больше $t_{кр}$ гипотеза отвергается, если же $t < t_{кр}$, то гипотеза H_0 принимается.

4. Если дана выборка объема n из бесконечной генеральной совокупности, то доверительные границы с доверительной вероятностью (1—q) определяются по формуле

$$M^B - t_{n-1} \frac{s}{\sqrt{n}} < M^\Gamma < M^B + t_{n-1} \frac{s}{\sqrt{n}}$$

где M^Γ — среднее арифметическое генеральной совокупности, M^B — среднее арифметическое выборки, q — уровень значимости, (1—q) — доверительная вероятность, s — несмещенная оценка дисперсии, которая рассчитывается по выборке, t_{n-1} — коэффициент, определяемый из таблицы распределения Стьюдента для столбца, соответствующего q и строки, соответствующей (n — 1) степени свободы.

5. *Пример 30.* Рассмотрим данные социально-демографического исследования молодоженов[27], подавших заявле-

[27] *ЧуйкоЛ. В.* Браки и разводы. М., 1975, с. 88 (рассчитано по табл. 27).

ние о вступлении в брак в Киевский Дворец бракосочетаний в 1970 г.:

Число супружеских пар	Зарплата (или стипендия) жениха (руб.)	Средняя зарплата невесты (руб.)	Несмещенная оценка среднеквадратического отклонения зарплаты невесты (руб.)
132	до 50	62,8	3,4
144	50—100	81,6	2,7
461	100—150	84,9	3,3

Проверим, связана ли зарплата жениха с зарплатой невесты. Для этого определим сначала значимо ли различаются на 1 %-ном уровне зарплаты невест для двух групп женихов:

с зарплатой до 50 и от 50 до 100 руб.

Итак, $n_1 = 132$, $M^B_1 = 62,8$, $s^B_1 = 3,4$,

$n_2 = 144$, $M^B_2 = 81,6$, $s^B_2 = 2,7$,

Тогда $\frac{\left(s_1^B\right)^2}{n_1} = 0,0875, \frac{\left(s_2^B\right)^2}{n_2} = 0,0506, t = \frac{18,8}{\sqrt{0,138}} = 50,6, \nu \approx 252$

Как видим, полученное значение t намного превышает требуемое для 1%-го уровня (2,58), т.е. связь есть.

Упражнение 91. По приведенным в примере данным проверить на 1%-ном уровне значимость различий средней зарплаты невест для двух групп женихов: с зарплатой 50- 100 руб. и 100—150 руб.

Ответ: $t = 12,1$; $\nu = 290$, различие значимо.

7. Значимость различий дисперсии

Из двух бесконечных нормально распределенных генеральных совокупностей (предположение о нормальности распределений здесь существенно, если исследователь сомневается в его верности, следует использовать другие методы[28]) извлечены независимые выборки объема n_1 и n_2. Требуется определить, равны ли дисперсии генеральных совокупностей.

1. $H_0 : \left(\sigma_1^\Gamma\right)^2 = \left(\sigma_2^\Gamma\right)^2$
   $H_1 : \left(\sigma_1^\Gamma\right)^2 \neq \left(\sigma_2^\Gamma\right)^2$

[197]

[28] *Закс Л.* Статистическое оценивание. М., 1976, с. 242, 243. *Гласс Дж., Стэнли Дж.* Статистические методы в педагогике и психологии. М., 1976.

2. Обозначим через $\left(\sigma_1^B\right)^2$ большую из выборочных дисперсий. Для проверки гипотезы H_0 против H_1 рассчитывается отношение выборочных дисперсий или оценок (дело в том, что $\frac{\left(\sigma_1^B\right)^2}{\left(\sigma_2^B\right)^2} = \frac{\left(s_1^2\right)}{\left(s_2^2\right)}$):

$$F = \frac{s_1^2}{s_2^2} \quad \text{(V,7,1)}$$

Jтношение дисперсий F имеет так называемое F-распределение с (n_1 — 1) и (n_2 — 1) степенями свободы (если верна гипотеза H_0).

3. Критические точки для заданного исследования уровня значимости определяются так: верхняя критическая точка F_B – по специальным таблицам (см. Приложение 3, табл. Л), а нижняя F_H из соотношения:

$$F = \frac{1}{F_B} \quad \text{(V,7,2)}$$

Гипотеза принимается, если рассчитанное по формуле (V,7,1) значение F лежит между F_H и F_B, *т.е.* $F_H < F < F_B$. Если же $F < F_H$ или $F > F_B$, то гипотеза отвергается на заданном исследователем уровне значимости.

4. Доверительный интервал с использованием найденных выше критических точек определяется по формуле:

$$\frac{s_1^2}{s_2^2} \cdot F_H < \frac{\left(\sigma_1^\Gamma\right)^2}{\left(\sigma_2^\Gamma\right)^2} < \frac{s_1^2}{s_2^2} \cdot F_B$$

Пример 81. Пусть $n_1 = 31$, $s_1^2 = 16$, $n_2 = 21$, $s_2^2 = 25$. Отношение большей оценки к меньшей равно: $\frac{25}{16} = 1{,}56$. F_B по таблице приблизительно равно 2,0; $F_H = 0{,}5$. Гипотеза H_0 принимается. $1{,}56 \cdot 0{,}5 < \frac{\left(\sigma_1^\Gamma\right)^2}{\left(\sigma_2^\Gamma\right)^2} < 1{,}56 \cdot 2$,т.е. $0{,}78\left(\sigma_1^\Gamma\right)^2 < \left(\sigma_2^\Gamma\right)^2 < 3{,}12\left(\sigma_1^\Gamma\right)^2$ - это доверительный интервал для большей дисперсии.

Упражнение 92.

Пусть $n_1 = 60$, $s_1^2 = 10$, n_2, == 140, s_2^2 == 5. Определить значимость различий на уровне 5- и 95% -и доверительный интервал для отношения генеральных дисперсий. Ответ: $F = 2$, $F_B = 1{,}43$, $F_H = 0{,}70$, д*оверительный* интервал

(1,40; 2,86).

8. Значимость коэффициентов корреляции *r*, ρ, τ и коэффициентов, основанных на χ^2

А. Коэффициент r

1. Требуется проверить значимость r, т.е. может ли при данном значении выборочного коэффициента r^B быть равным нулю коэффициент корреляции r^Γ для генеральной совокупности.

1. H_0: $r^\Gamma=0$

2. H_1: $r^\Gamma \neq 0$

2. Статистика *t,* рассчитываемая по формуле (V,8,1) имеет *t*-распределение с *(n*—2) степенями свободы:

$$t = \frac{r^B \sqrt{n-2}}{\sqrt{1-(r^B)^2}} \text{(V,8,1)}$$

3. Критические точки[29] определяются по таблице И Приложения 3 для заданного уровня значимости *q,* при $|t| \geq t_{кр}$ гипотеза $H_0 : r^\Gamma = 0$ отклоняется на уровне значимости *q*.

4. Для построения доверительных интервалов выборочное значение r^B подвергается так называемому преобразованию Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} \text{ (V8.2)}$$

Дело в том, что r^B имеет нормальное распределение лишь в случае, когда $r^\Gamma = 0$, а вот z^B — при любых значениях z^Γ. Поэтому, рассчитав z^B по полученному в выборке значению r^B, строим доверительный интервал для r^Γ (*z* определяется по таблице нормального распределения А Приложения 3, например для 95% доверительного интервала *z* =1,96):

$$z^B - z\frac{1}{\sqrt{n-3}} < z^\Gamma < z^B + z\frac{1}{\sqrt{n-3}}$$

Получив нижнее и верхнее значения z, рассчитываем значения $r_{нижн}$ И $r_{верхн}$ по формуле

$$r = \frac{e^{2z}-1}{e^{2z}+1} \text{ (V.8.3)}$$

Преобразование (V,8,3) можно также осуществить по табл. К Приложения 3.

[199]

[29] Значимость *r* можно оценить также непосредственно по таблице Е Приложения 3 (без проведения каких-либо расчетов).

5. Рассмотрим следующий *пример* 32. В исследовании инженеров ленинградских проектноконструкторских организаций были получены данные, характеризующие связь удовлетворенностей профессией и работой[30]. Коэффициент корреляции, рассчитанный по оценкам 89 руководителей групп, между удовлетворенностью работой и удовлетворенностью профессией равен 0,23. Проверим значимость коэффициента на 1%-ном уровне и построим соответствующий (т.е. 99%-й) доверительный интервал.

Итак, $n = 89$, $r^B = 0{,}23$. По формуле (V,8,1) получаем:

$$t = \frac{0{,}23\sqrt{89-2}}{\sqrt{1-(0{,}23)^2}} \approx 2{,}2$$

По табл. И Приложения 3 находим, что для n — 2 = 87 степеней свободы 1%-я критическая точка равна приблизительно 2,64 (в таблице не приведены критические значения для 87 степеней свободы, но приведены для 60, равные 2,66 и для 120 степеней свободы — 2,62, т.е. искомое значение критической точки лежит приблизительно посредине между 2,62 и 2,66). Таким образом, хотя в цитированной книге, откуда взят этот пример, указано, что коэффициент значим на уровне 1%, он значим лишь на 5%-ном уровне (как видно из таблицы, критическое значение для уровня 5% $t_{кр}$ =1,99). Такой же результат дает использование таблицы Е Приложения 3.

Построим теперь 99%-и доверительный интервал. По формуле (V,8,2) получаем:

$$z^B = \frac{1}{2}\ln\frac{1+0{,}23}{1-0{,}23} \approx 0{,}234$$

По таблице нормального распределения находим, что для 1%-го уровня значимости z=2,58. Тогда нижняя граница для $z^Г$ равна $0{,}234 - 2{,}58\frac{1}{\sqrt{89-3}} = -0{,}04$ Верхняя граница равна $0{,}234 + 2{,}58\frac{1}{\sqrt{89-3}} = 0{,}513$.

Таким образом, —0,045 < $z^Г$ < 0,513. Но это доверительные границы для z. Теперь необходимо опять вернуться к коэффициентам корреляции. По формуле (V,8,3) получаем для

[200]

[30] Социально-психологический портрет инженера. М., 1977, табл. 11 на с. 149.

нижней границы:

$$\frac{e^{-2\cdot 0,045}-1}{e^{-2\cdot 0,045}+1}=-0,04$$

Для верхней границы аналогичные расчеты дают 0,47.

Вместо преобразований по формулам (V,8,2) и (V,8,3) приблизительно тот же результат можно получить по таблице К Приложения 3. Например, для преобразования 0,23 находим строку 0,2 и столбец 3 — на их пересечении стоит *z,* соответствующий $r = 0,23$, а именно 0,2342. А чтобы преобразовать $z = 0,513$, находим внутри таблицы наиболее близкое к нему число (это 0,5101). Оно стоит в строке 0,4 и столбце 7, следовательно, $r = 0,47$. Итак, 99% доверительный интервал равен:

$-0,04 < r^{\Gamma} < 0,47$,

По таблице Ж Приложения 3 можно найти, не проводя расчетов, приближенно 95%-й интервал: $0,03 < r^{\Gamma} < 0,42$.

6. *Упражнение 93.* В той же работе[31] указывается, что коэффициент корреляции, полученный для 52 главных инженеров проекта и равный 0,51, значим на уровне 1%. Проверить, так ли это, и построить доверительный интервал. Ответ: коэффициент значим *(*$t = 4,19$); $0,19 < r^{\Gamma} < 0,73$.

Б. Коэффициент ранговой корреляции ρ

1. H_0: $\rho^{\Gamma}=0$

2. H_1: $\rho^{\Gamma}\neq 0$

2. Если H_0 верна и $n > 10$ (при $n \leq 10$ значимость ρ определяется другим способом по специальной таблице[32], то значение *t,* рассчитанное по формуле (V,8,4), имеет *t* — распределение Стьюдента с *(*n — 2) степенями свободы:

$$t=\frac{\rho^{B}}{\sqrt{\left[1-(\rho^{B})^{2}\right]\frac{1}{n-2}}} \quad \text{(V,8,4)}$$

3. Критические точки определяются по табл. И Приложения 3. Значимость ρ можно определить также непосредст-

[201]

[31] Социально-психологический портрет инженера. М., 1977, с. 149.

[32] *Кендэл М.* Ранговые корреляции. М., 1975, с. 69, 188—191.

венно по значению ρ (без расчета *t)* по таблице В Приложения 3.

4. Доверительные интервалы для коэффициента ρ не рассчитываются, так как, оказывается, получить выборочное распределение для ρ^B в случае, когда $\rho^{\Gamma} \neq 0$, очень сложно[33].

Пример 33. В примере 19 был рассчитан коэффициент связи между положительными ответами на вопросы «интересная работа» и «образование соответствует работе» для 14 групп рабочих. Оказалось, что ρ^B=0,345. Определим на 5%-ном уровне значимость этого коэффициента.

$$t = \frac{0{,}345}{\sqrt{\left[1-(0{,}345)^2\right]\cdot\frac{1}{12}}} \approx 1.27$$

Поскольку это меньше критического значения 2,23 (для 10 степеней свободы, так как в таблице не приведены значения для 12), коэффициент незначим, хотя в цитированной работе он интерпретируется как значимый.

Упражнение 94. Определить, будет ли значим коэффициент, если он рассчитан для 42 групп. Ответ: Да *(t* = 2,32).

В. Коэффициент ранговой корреляции τ

Вопрос о существенности коэффициента т мы рассматривали ранее (§ 6, гл. II), там же показано, каким образом определять значимость τ при $n \leq 10$ (поскольку для проверки существенности используется S, мы сочли целесообразным рассмотреть этот вопрос сразу после введения S). Пусть $n > 10$.

1. H_0: $\tau^{\Gamma}=0$

2. H_1: $\tau^{\Gamma} \neq 0$

2. Если H_0 верна, величина z имеет нормальное распределение:

$S^* = S - 1$, если $S>0$

$z = \frac{S^*}{\sigma}$, где S*=0, если S=0; (VI,8,5)

$S^* = S + 1$, если $S<0$

$\sigma = \sqrt{\frac{1}{18} n(n-1)(2n+5)},$

[202]

[33] *Кендэл М.* Ранговые корреляции. М., 1975, с. 102. См. также *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи. М., 1973, с. 637—644.

если нет объединенных раногов и

$$\sigma^2 = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_r t_r(t_r-1)(2t_r+5) - \sum_s u_s(u_s-1)(2u_s+5) \right] +$$

$$+ \frac{1}{9n(n-1)(n-2)} \cdot \left[\sum_r t_r(t_r-1)(t_r-2) \right] \cdot$$

$$\cdot \left[\sum_s u_s(u_s-1)(u_s-2) \right] + \frac{1}{2n(n-1)} \cdot \left[\sum_r t_r(t_r-1) \right] \cdot \left[\sum_s u_s(u_s-1) \right]$$

если есть объединенные ранги; σ, разумеется, равна корню квадратному из приведенного выражения; t_r и u_s — число объединенных рангов в r-м объединении по X и s-м объединении по Y соответственно.

3. Критические точки определяются по таблице нормального распределения, H_0 отвергается при $|z| > z_{кр}$. При отсутствии объединенных рангов значимость τ можно определить по таблице Д Приложения 3 (без расчета г).

4. Доверительные интервалы для τ не определяются из тех же соображений, что и доверительные интервалы для ρ.

5. *Пример 34.* Пусть $n = 20$, $S = 52$ ($\tau = 0{,}27$). Определим значимость τ на уровне 5%. Так как S положительно, S^* равно 51.

$$z = \frac{51}{\sqrt{\frac{1}{18} \cdot 20 \cdot 19 \cdot 45}} \approx 1{,}65$$

Так как для 5%-ного уровня критическое значение равно 1,96 (табл А Приложения 3), гипотеза H_0 принимается, коэффициент незначим.

Упражнение 95. $S = 33$ ($\tau = 0{,}36$), $n = 14$. Найти, значим ли коэффициент на уровне 5%. Ответ: незначим, z=1,75.

Г. Коэффициенты, основанные на χ^2

Как уже указывалось, существенность их проверяется с помощью χ^2. если значим χ^2, то значим и рассчитанный с его помощью коэффициент. Поэтому при расчетах, которые для таблиц $k \times l$, как правило, производятся на ЭВМ, желательно выпечатывать не только значение коэффициента,

[203]

но и значение χ^2. Если это не сделано, то χ^2 можно, разумеется, легко найти, преобразовав формулы для расчета коэффициентов. Например, преобразовав формулу для расчета коэффициента Чупрова *T,* получим:

$$\chi^2 = T^2\sqrt{n(k-1)(l-1)}$$

1. Нулевая гипотеза H_0 состоит в том, что $N_{ij} = N(x_i)N(y_j)\frac{1}{N}$ для всех *i* и *j*. Гипотеза H_1 заключается в том, что найдется хотя бы одна пара *i* и *j* такая, что $N_{ij} \neq N(x_i)N(y_j)\frac{1}{N}$. Критическая точка определяется для заданного исследователем уровня значимости *q* и для *(k*—1)·*(l*—1) степеней свободы по таблице Б Приложения 3.

Доверительные интервалы для χ^2 не вычисляются. *Пример 35.* Значение χ^2, рассчитанное для таблицы 21 (гл. II, §2), равно 92,2. Поскольку *(k*—1)·*(l*—1)=10, χ_0^2 для 1%-го уровня значимости равно 23,21. Следовательно, χ^2 и все рассчитанные на его основе коэффициенты значимы.

Упражнение 96. Для таблицы 5×8 было получено значение, равное 45,4. Значимо ли это значение на уровне 5%? Ответ: да.

9. Значимость различий r_1 и r_2

Из двух бесконечных генеральных совокупностей извлечены выборки объема n_1 и n_2 и для некоторых признаков *X* и *Y* в каждой из выборок рассчитаны выборочные коэффициенты корреляции r_1^B и r_2^B.

1. $H_0 : r_1^\Gamma = r_2^\Gamma$

2. $H_1 : r_1^\Gamma \neq r_2^\Gamma$

2. Для проверки гипотезы H_0 применяется z-преобразование Фишера — см. формулу (V,8,2). Вычисляем r:

$$z = \frac{z_1^B - z_2^B}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \quad \text{(V,9,1)}$$

Можно доказать, что *z* имеет нормальное распределение с нулевым средним и единичной дисперсией, если верна гипотеза H_0.

[204]

3. Критические точки по заданному уровню значимости *q* определяются по таблице А Приложения 3.

4. Найденные в п. 3 критические точки могут быть использованы для построения доверительных интервалов с доверительной вероятностью 1 — *q:*

$$\left(z_1^B - z_2^B\right) - z\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} < \left(z_1^\Gamma - z_2^\Gamma\right) < \left(z_1^B - z_2^B\right) + z\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

После того, как найдены критические точки для z, они преобразуются в критические точки для *r* по формуле (V,8,3) или с помощью таблицы К Приложения 3.

5. *Пример 36.* В исследовании влияния престижа профессий среди школьников на привлекательность профессии[34] была выдвинута гипотеза, что престиж оказывает большее влияние на привлекательность профессии для школьников из семей интеллигенции, чем для школьников из семей рабочих. Для проверки этого предположения было проведено репрезентативное для г. Киева исследование школьников 10-х классов, в ходе которого для 43 профессий были получены оценки престижа и привлекательности их для школьников. Связь престижа и привлекательности характеризовалась коэффициентом корреляции этих оценок. Оказалось, что этот коэффициент для школьников из семей интеллигентов равен 0,94, а из семей рабочих 0,82. Дают ли эти данные основание заключить, что гипотеза подтвердилась? Проверим значимость различий: $n_1 = n_2 = 43$, $r^B_1 = 0{,}94$, $r^B_2 = 0{,}82$. Чтобы воспользоваться формулой (V,9,1), проводим сначала *z*-преобразование Фишера по формуле (V,8,2) или по таблице К Приложения 3: $z_1 = 1{,}738$, $z_2 == 1{,}157$. Тогда получим:

$$z = \frac{1{,}738 - 1{,}157}{\sqrt{\frac{1}{43 - 3} + \frac{1}{43 - 3}}} \approx 2{,}60.$$

Поскольку полученное значение выше, чем 2,58, можно утверждать, что различия значимы на уровне 1%, и гипотеза

[205]

[34] *Черноволенко В. Ф., Оссовский В. Л., Паниотто В. И.* Престиж профессий и проблемы социально-профессиональной ориентации молодежи. Киев, 1979, с. 146, 147.

исследователей, следовательно, получила эмпирическое подтверждение.

Для этого же уровня значимости (т.е. для 99% доверительной вероятности) найдем доверительный интервал:

$$(1{,}738-1{,}157)-2{,}58\sqrt{\frac{2}{40}}<z_1^{\Gamma}-z_2^{\Gamma}<(1{,}738-1{,}157)+2{,}58\sqrt{\frac{2}{40}}$$

или $0{,}004 < z_1^{\Gamma} - z_2^{\Gamma} < 1{,}158$

Переводя z в r по формуле (V,8,3) или по таблице К Приложения 3, получаем: $0{,}004 < z_1^{\Gamma} - z_2^{\Gamma} < 0{,}820$

6. *Упражнение 97.* В примере и упражнении § 8 А приведены результаты исследования[35], показавшего, что коэффициент корреляции между удовлетворенностями работой и профессией 89 руководителей групп равен 0,23, а этот же коэффициент, рассчитанный по оценкам 52 главных инженеров проекта, равен 0,51. Достаточны ли эти различия, чтобы утверждать, что более высокое должностное положение позволяет полнее реализовать профессиональные ожидания, значимо ли различие полученных коэффициентов корреляции? Проверить на 5%-м уровне значимости и построить для разности коэффициентов доверительный интервал.

Ответ: различие незначимо (z = 1,83), $-0{,}022 < \left| r_1^{\Gamma} - r_2^{\Gamma} \right| < 0{,}592$. Отметим связь гипотез с доверительными интервалами: если H_0 принимается, то доверительный интервал содержит 0; если H_0 отвергается — то не содержит. Как видим, в данном случае доверительный интервал содержит 0.

[206]

[35] Социально-психологический портрет инженера. М., 1977, табл. 11, с. 149

Глава VI

КЛАССИФИКАЦИЯ ОБЪЕКТОВ (ТАКСОНОМИЯ), КЛАССИФИКАЦИЯ ПРИЗНАКОВ (ФАКТОРНЫЙ АНАЛИЗ) И НЕКОТОРЫЕ ДРУГИЕ МЕТОДЫ АНАЛИЗА ИНФОРМАЦИИ

Кроме описанных, в социологических исследованиях используются и другие методы анализа информации, обзору которых и посвящена настоящая глава.

В §9 главы 1, описывая матрицу данных (табл. 1), мы дали эмпирическую информацию компактной, удобной для анализа. Рассмотренный нами ранее путь заключается в расчете характеристик, описывающих распределение опрошенных по каждому признаку. Например, использование средних позволяет «свернуть» матрицу данных в одну строку, состоящую из средних характеристик всего массива по каждому из изучаемых признаков. Такое представление, однако, во-первых, эффективно при достаточной однородности объектов по изучаемым признакам, во-вторых, не решает полностью задачу конденсации информации при большом числе признаков. Рассматриваемые ниже методы позволяют, с одной стороны, «сжать» матрицу данных, классифицируя опрошенных[1] и объединяя их в небольшое число однородных групп (таксономия), с другой стороны, позволяют объединить признаки в небольшое число групп (факторный анализ).

Таксономия. В качестве синонимов для обозначения этой группы методов используют также термины «кластерный анализ», «авто классификация» или (более широко) говорят об использовании методов «распознавания образов». Пусть матрица данных включает характеристики N объектов по двум количественным признакам (например, стаж работы и зарплата). Откладывая признаки по осям координат, мы можем изобразить все объекты на плоскости в виде N точек: абсцисса – значение стажа, ордината – значение зарплаты данного объекта. В этом случае говорят, что N объектов

[207]

[1] Поскольку в матрице данных могут быть не только индивиды, но и бригады, предприятия, населенные пункты и т. п., мы будем далее говорить об «объектах», а не об «опрошенных».

расположены в двухмерном признаковом пространстве; (по сути, это один из способов изображения двухмерного распределения признаков). Как видно из рисунка, все объекты можно разбить на три группы таким образом, что объекты внутри групп близки между собой (это означает, что они имеют близкие характеристики и по Х и по У), а объекты из разных групп – далеки.

Множество близких между собой точек называется *таксоном* и при интерпретации результатов рассматривается

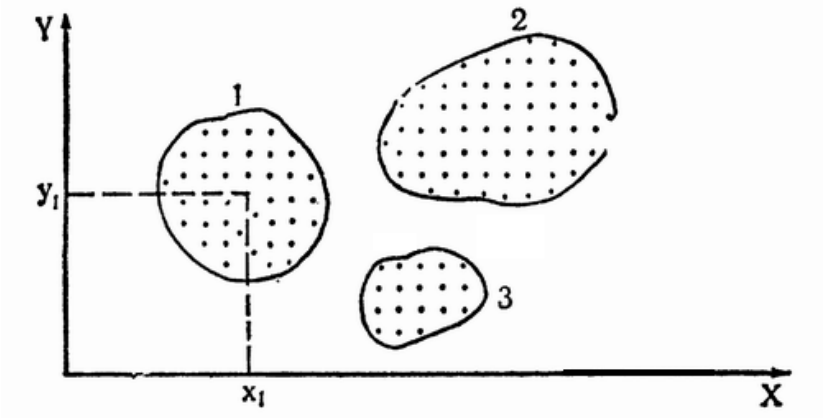


Рис. 28. Изображение объектов в пространстве двух признаков (1, 2, 3 – таксоны)

как некоторый социальный тип. Если имеется *k* признаков, то говорят, что объекты расположены в *k*-мерном| признаковом пространстве, Если признаков более чем два, то точки уже невозможно изобразить на плоскости. В этом случае группировку можно осуществить с помощью формальных методов, которые и называются методами таксономии.

Результатом работы алгоритмов таксономии обычно является разбиение множества объектов на группы (таксоны) в пространстве признаков, заданных исследователем, а также расчет некоторых обобщенных характеристик каждого из таксонов (центр таксона, средние, меры вариации). Существуют алгоритмы, позволяющие проводить классификацию не только в пространстве признаков, измеренных с помощью метрических шкал, но и для шкал номинальных и порядковых.

В качестве примера рассмотрим применение таксономии для классификации сельских мигрантов[2]. Задача заключалась в проверке гипотезы о том, что направление мигра-

[208]

[2] Распознавание образов в социологических исследованиях. Новосибирск, 1968.

ции зависит от пола, возраста, образования, семейного положения, числа детей и типа профессии, который характеризовался числом месяцев обучения. Выделение типов мигрантов дало возможность разрабатывать мероприятия, направленные на регулирование миграции дифференцированно и, следовательно, более эффективно. С помощью методов

Таблица 40

Характеристика групп, полученных методом таксономии

№№ группы	Общая характеристика	Средние показатели по группе			
		Возраст лет	Число детей	Общее образование, лет	Специальное образование, месяцев
1	Семейные мужчины и женщины	30,6	1,7	5,4	2,0
2	Неженатые молодые мужчины	22,1	0,2	6,5	2,4
3	Молодые девушки и незамужние женщины	19,7	0,1	7,1	0,8
4	Престарелые женщины без мужей	61,0	0,9	1,3	0,1
5	Одинокие женщины среднего возраста без специальности	32,1	0,1	1,4	0,9
6	Одинокие женщины специалисты	24,8	0,2	11,6	41,0

таксономии все мигранты, или «точки» в шестимерном пространстве из перечисленных признаков, были разбиты на 6 групп (таксонов): в один таксон попадали лица, близкие по приведены в табл. 40 (совпадение числа признаков пространства с числом полученных групп, разумеется, случайно).

Чтобы проверить, действительно ли эти группы представляют разные типы, сравним характер их миграции (табл. 41). Мы видим, что группы существенно различны. Например, первая – семейные мигранты с детьми - дает наибольшую долю движения в пределах сельской местности, к ней приближается 4-я группа - «бабушки», но «бабушки» явно отличаются от группы! по направлению миграции. С помощью методов таксономии решались также задачи блокировки профессий, выявления групп рабочих по степени устойчивости на предприятии и др.,

[209]

Другим направлением конденсации информации является *факторный анализ* признаков. Как уже неоднократно отмечалось, индивиды обладают самыми разнообразными признаками, которые не являются независимыми. Связи между ними, как мы видели, изучаются с помощью методов корреляционного анализа. Можно предположить, что некоторые признаки образуют группы, каждая из которых

Таблица 41

Распределение мигрантов, вошедших в разные группы, по характеру движения между городом и деревней

Характер движения	Номера групп					
	1	2	3	4	5	6
1.Из крупного или среднего города в село	7,3	0,8	0	1,2	0	3,0
2. Из малого города в село	8,0	6,6	3,2	10,6	0	21,2
3. Из села в село	47,4	14,6	13,8	41,7	22,8	15,2
4. Из села в малый город	16,5	47,0	35,0	4,1	31,0	22,7
5. Из села в средний или крупный город	20,8	31,0	48,0	42,4	46,2	37,9

отражает определенный аспект сложного явления. При анализе системы признаков мы сталкиваемся не с классификацией объектов, а с классификацией признаков, т.е. с выявлением групп признаков, имеющих сходный характер изменения при переходе от одного объекта к другому. В частности, ставится задача найти максимально взаимосвязанные группы признаков. Выделяемые группы – это новые, комплексные переменные, называемые *факторами.*

Факторный анализ позволяет не только выделить группы наиболее взаимосвязанных признаков, но и отделить несущественные признаки от существенных, оценить их информативность.

Обоснованная замена большого числа признаков, описывающих объекты наблюдения, меньшим числом комплексных характеристик (факторов) составляет сущность факторного анализа.

Подчеркнем, что факторы не сводятся к некоторым, пусть главным, основным признакам исходного набора, Каждый фактор - это группа взаимосвязанных признаков из упомянутого набора, и вся совокупность входящих в него признаков определяет содержательную интерпретацию этого фактора.

[210]

Выделение групп признаков, подобно выделению таксонов, означает «конденсацию» информации, построение более простого описания, которое помогает вскрыть логическую структуру изучаемого явления, выделить наиболее характерные связи в системе признаков, проверить гипотезы о взаимосвязях, выдвинуть новые н т.д.

Попутно отметим, что выделение факторов упрощает решение задачи многомерной классификации объектов наблюдения, т.е. группировки объектов со сходными значениями признаков (задача таксономии). Здесь факторный анализ выступает как предклассификация, предварительный этап классификации объектов, Переход к небольшому числу комплексных переменных (факторов) упрощает применение графического анализа, интерпретацию результатов,

Рассмотрим конкретный пример[3] применения факторного анализа к изучению природы стимулирующего воздействия на трудовую деятельность,

Общая схема стимулирующего акта может быть представлена следующим образом: создаются условия для реализации целей, формируется сознательная ориентация работников на выполнение цели. Необходимым условием достижения целей являются высокие показатели в работе (объективное отношение к труду, или фактическое поведение в сфере трудовой деятельности). Фактическое поведение фиксировалось как выполнение норм выработки, качество работы, дисциплинированность, участие в рационализации и изобретательстве.

При изучении субъективного отношения к труду, или ориентации на трудовую деятельность, рассматривались: отношение к работе, к специальности, к различным элементам производственной ситуации,

В исследовании фиксировались также демографические и функциональные признаки работников, которые можно рассматривать как референты социальных условий жизнедеятельности, как признаки, свидетельствующие о диапазоне реальных возможностей той или иной группы работников в определенной системе общественных отношений.

Таким образом, члены производственного коллектива являются носителями ряда признаков, а именно: демогра-

[211]

[3] Исследование проводилось социологическим отделом Одесского отделения ИЭ АН УССР в 1971-75 гг. под руководством И. М. Поповой. Разработка методики сбора, обработки и анализа информации осуществлялась В. С. Максименко.

фические и функциональные признаки, оценки-ориентации разного рода, различные качества работника и т.д.,

Естественно предположить (и это в известной мере заложено в приведенной выше интуитивной априорной классификации, вытекающей из предварительного теоретического анализа), что некоторые признаки могут быть объединены в группы, т.е. возникает задача группировки признаков. Особый интерес при этом представляет характер связи различных групп признаков с конечным результатом стимулирования – фактическим поведением, конечно, если образуется группа признаков, описывающая это поведение,

Для анализа была выбрана следующая система признаков работающих:

1) квалификация;
2) стаж работы на заводе;
3) стаж работы по данной специальности;
4) образование;
5) возраст;
6) величина заработной платы;
7) выполнение норм выработки;
8) состояние трудовой дисциплины;
9) качество работы;
10) участие в рационализации и изобретательстве;
11) удовлетворенность работой (предприятием);
12) оценка степени физической нагрузки (тяжела ли работа физически?);
13) удовлетворенность содержанием труда интересна ли работа?);
14) оценка организации труда (простои, "штурмовщина");
15) удовлетворенность заработной платой;
16) удовлетворенность отношениями с администрацией;
17) мнение о справедливости распределения премий;
18) удовлетворенность специальностью.

Собор признаков, с одной стороны, диктовался стремлением учесть социально-демографические характеристики работников, объективное отношение к труду и субъективное (удовлетворенности оценки как интегральные – работой в целом, специальностью, так и частные – отдельными элементами рабочей ситуации). А с другой стороны, лимитировался техническими возможностями расчета корреляций между признаками, число которых М, как известно, является квадратичной функцией числа признаков n: $M = \frac{n(n-1)}{2}$

Информация о системе отобранных признаков содержится в матрице корреляций, которая была построена на основе коэффициента Чупрова[4].

Рассмотрим основные результаты применения факторно-

[4] Обсчет информации осуществлялся в Институте проблем управления АН СССР И. Б. Мучником и Н. Е. Киселевой по алгоритму, изложенному в статье Э. Бравермана, А. Дорофеева, М. Луганского, И. Мучника «Методы диагонализации матриц связи» (Проблемы расширения возможностей автоматов. Труды Ин-та проблем управления, вып. 1, 1973).

го анализа. При выделении двух факторов в одну группу попадают все признаки, характеризующие субъективное отношение к трудовой деятельности (11 - 13), во вторую - остальные, описывающие и социально-демографические характеристики работников и объективное отношение к трудовой деятельности.

Все удовлетворенности-оценки (первый фактор) тесно взаимосвязаны, хотя они связаны и с социально-демографическими признаками работающих, и с объективным отношением к труду (второй фактор), эта связь меньше, чем взаимосвязь; корреляция факторов 0,269, а факторные нагрузки[5], описывающие корреляцию признаков, характеризующих субъективное отношение к труду с фактором, заключены между 0,60I и 0,417. Можно предположить, что эмпирический материал свидетельствует об относительной самостоятельности сферы сознания. Отметим, что социально-демографические признаки, попадающие во второй фактор, в большей степени связаны с объективным, чем субъективным отношением к трудовой деятельности.

В группе признаков, описывающих субъективное отношение к труду, максимальная факторная нагрузка у признака «отношение к специальности» (0,601), далее идут «отношение к содержанию труда» (0,524), «отношение к работе в целом» (0,520) и т.д.

Во второй группе признаков на первое место по величине факторной нагрузки выходит возраст (0,582), на второе – квалификация (0,551), на третье – качество работы (0,520) – первый из признаков, описывающих объективное отношение к труду – и т.д.

Обратим внимание на то, что внутри второго фактора социально-демографические признаки не локализованы, а чередуются с признаками, описывающими объективное отношение к труду.

При выделении трех факторов образуются группы, описывающие:

1) f_1 социальные условия жизнедеятельности (1 - 6);

2) f_2 объективное отношение к труду, показатели в работе (7 - 10);

3) f_3 субъективное отношение к труду, ориентацию на трудовую деятельность (11 - 18).

[213]

[5] Коэффициенты, характеризующие связь признаков с фактором; их можно интерпретировать как коэффициенты корреляции фактора с признаками.

Таким образом, как бы распадается на две части группа признаков, составлявшая ранее второй фактор. Теперь все факторы состоят из *сходных* признаков и тем самым могут быть естественно интерпретированы. Оказывается, что максимально взаимосвязаны f_1 и f_2, минимально f_2 и f_3, т.е. подтверждаются и детализируются выводы, сделанные ранее при рассмотрении двух факторов.

Внутри признаков, описывающих объективное отношение к труду, максимальная факторная нагрузка у такого признака, как качество работы, на втором месте - выполнение норм, далее идут дисциплинированность, участие в рационализации и изобретательстве. (Эта последовательность сходна с ранее полученной. Заметим, что она сохраняется в дальнейшем при переходе к большему числу факторов).

Как мы видим, два последних признака относительно менее информативны. Это связано с тем, что: 1) практически все работники дисциплинированы и 2) большая часть их в рационализации не участвует.

В группе социально-демографических признаков по-прежнему на первом месте – возраст, на последнем – образование, а стаж по специальности «опережает» стаж на заводе. Последовательность признаков – возраст, стаж по специальности, квалификация, заработная плата, стаж на заводе, образование – сохраняется и при переходе и большему числу факторов.

То обстоятельство, что образование менее тесно связано с фактором, чем стаж, по-видимому, отражает специфику объекта – судоремонтные предприятия. Как было выяснено в ходе исследований, для судоремонтных профессий при прочих равных условиях - стаж в большей мере определяет результаты трудовой деятельности, а также квалификацию рабочих, чем образование.

То, что стаж по специальности в большей степени связан с фактором, чем стаж работы на предприятии, можно также рассматривать как следствие специфики судоремонта, где рабочим приходится сталкиваться с самыми разнообразными типами судов, и профессиональные навыки, референтом которых является стаж по специальности, в результатах трудовой деятельности играют более важную роль, чем адаптация к условиям данного предприятия, референтом которой можно считать стаж работы на предприятии.

В случае четырех факторов группа признаков субъективного отношения распадается на две. В первую входят 4 при-

[214]

знака, находившихся на более высоких местах: удовлетворенность специальностью, содержанием труда, предприятием в целом, отношением с администрацией. Во вторую - остальные. Внутри каждой из этих групп последовательность признаков практически такая же, как и в исходной группе, из которой они образовались.

С дальнейшим увеличением числа факторов (5, 6...) результаты становятся менее надежными, и мы не станем их приводить.

В заключение рассмотрим основные содержательные результаты применения факторного анализа к проблеме изучения стимулирующего воздействия на трудовую деятельность.

Все выделенные признаки оказываются практически взаимосвязанными, это можно рассматривать как свидетельство сложной природы стимулирующего воздействия.

Предложенная исходная группировка признаков целесообразна. Различные группы признаков характеризуют разные, относительно самостоятельные уровни регулирования трудовой деятельности: связь между признаками одного и того же уровня более тесная, чем между признаками разных уровней.

При исследовании взаимосвязей различных групп признаков обращает на себя внимание относительная самостоятельность сферы сознания (оценки, удовлетворенности), ее относительная ограниченность от сферы фактического поведения и - в меньшей мере - от признаков, характеризующих условия жизнедеятельности.

Фактическое поведение в сфере трудовой деятельности в большей степени определяется социально-демографическими признаками, чем субъективным отношением и труду.

Выше рассмотрены результаты факторного анализа признаков, корреляции между которыми описывались с помощью коэффициента Чупрова. В дальнейшем факторному анализу была подвергнута матрица корреляции, построенная на основе коэффициента Крамера, теоретически более предпочтительного. В принципе результаты получились близкие. Отметим только, что такой признак, как заработная плата, который в первом случае попадал в группу социально-демографических, во втором вошел в группу признаков, описывающих фактическое поведение в сфере трудовой деятельности. При выделении трех и четырех факторов эта группа остается компактной; последовательность признаков - качество работы, выполнение норм выработки, участие в

[215]

рационализации и изобретательстве, заработная плата, дисциплинированность – сохраняется.

«Переход» признака заработная плата в фактор, описывающий фактическое поведение работников, предоставляется теоретически оправданным. Все остальные содержательные выводы подтверждаются.

Отметим, что логическая непротиворечивость и естественность интерпретации полученных результатов могут свидетельствовать о возможности применения коэффициентов для факторного анализа социальных признаков. Аналогом факторного анализа является *латентно-структурный анализ*[6].

Для углубленного изучения связей между признаками используется также причинный и дисперсионный анализ.

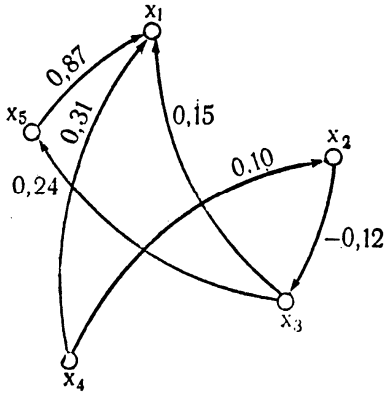


Рис. 29 Пример графа причинных связей для пяти признаков.

Причинный (или путевой) *анализ* используется для выявления *непосредственных* влияний одних признаков на другие[7]. Можно рассматривать путевой анализ, с одной стороны, как своеобразное развитие методов частной корреляции и, с другой – как поиск направленных мер, характеризующих влияние одного признака на другой. Результаты причинного анализа часто представляют как в виде ориентированного графа[8], вершинами которого обозначаются признаки, а дугами – влияния одних признаков на другие. Степень влияния *i*-го признака на *j*-й характеризуется специальным показателем[9] P_{ij}. На рис. 29 изобра-

[216]

[6] Статистические методы анализа информации в социологических исследованиях. М., 1979, гл. 14; Математические методы в современной буржуазной социологии. М., 1966.

[7] Математика в социологии. Моделирование и обработка информации. М., 1977. ч. 1; Статистические методы анализа информации в социологических исследованиях. М., 1979, гл. 15. Хейс Д. Причинный анализ в статистических исследованиях, М., 1981.

[8] Под ориентированным графом можно понимать некоторое множество точек (называемых вершинами графа), соединенных стрелками (называемых дугами).

[9] Интерпретация этих показателей напоминает интерпретацию коэффициентов регрессии, хотя и не совпадает с ней. (Хейс Д. Причинный анализ в статистических исследованиях, с. 37, 104).

жен пример графа причинных связей для пяти признаков. Видно, например, что 3-й признак влияет на 1-й признак как непосредственно (0,15), так и через 5-й признак, наибольшее непосредственное влияние на 1-й признак оказывает 5-й (0,87) и т.д.

Рассматривая свойства дисперсии (§4, гл. 1), мы вывели формулу (148), показывающую, что общая дисперсия состоит из межгрупповой и внутригрупповой. Это равенство лежит в основе другого подхода к изучению влияния одних признаков на другие, который называется *дисперсионным анализом.* При изучении влияния набора признаков на вариацию некоторого результирующего признака (например, различных стимулов на повышение производительности труда) дисперсионный анализ позволяет вычленить влияние каждого из признаков. Это дает возможность отойти от традиционного метода планирования эксперимента (поддерживание стабильными всех переменных и вычленение влияния одного признака на результирующий) и перейти к экспериментам, в которых одновременно изменяются все признаки. Поэтому дисперсионный анализ очень тесно связан с *планированием эксперимента*[10] (так называется один из разделов математической статистики) и может широко использоваться в планировании социологического исследования.

Чрезвычайно перспективным направлением статистики, развитым специально для нужд психологии, социальной психологии и социологии, является *многомерное шкалирование*[11]. Методы многомерного шкалирования позволяют продуцировать гипотезы о критериях, которыми пользуются индивиды для оценки различных объектов. Исходной информацией для использования этих методов являются эмпирические данные либо о ранжировании индивидами некоторого набора объектов (например, ранжировка профессий по привлекательности), либо о сходстве объектов между собой (например, респондентам предъявляют всевозможные сочетания по две профессии из всего множества профессий и просят оценить сходство каждой пары профессий с помощью балльной оценки).

По этим данным находят минимальное признаковое пространство (т.е. пространство с минимальным числом осей), в котором можно так разместить оцениваемые объекты, чтобы

[217]

[10] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М., 1979; Налимов В. В. Теория эксперимента. М., 1971.

[11] Клигер С. А., Косолапов М. С., Толстова Ю. Н. Шкалирование при сборе и анализе социологической информации. М., 1978.

сохранились такими же, как в эмпирических данных, порядок или показатели близости объектов. Как правило, размерность этого пространства невелика, и его можно наглядно изобразить. Если, например, это пространство размерности 2, то получим картину, аналогичную изображенной на рис. 28. Далее, каждая из осей интерпретируется как неявная шкала, которой пользуются респонденты для оценки объектов. Например, исследователь, обнаруживший, что проекции точек, изображающих профессии, на ось X легли в таком порядке - продавец-лоточник, водитель трамвая, корректор, монтажник радиоаппаратуры, техник связи, инженер, научный работник - может интерпретировать ось X как уровень образования, требуемый для данной профессии. Следующим шагом может быть проверка этого предположения в эмпирическом исследовании.

Важно отметить, что таким способом производится измерение по шкале, которую исследователь не задает априорно и, следовательно, не навязывает ее респонденту (обычный путь предполагает определение признаков, влияющих, например, на привлекательность профессии, до исследования, разработку шкал и включение их в анкету - при этом исследователь может пропустить важные для респондентов критерии оценки).

Отметим, наконец, две группы методов, отличающихся от изложенных не столько по возможностям, сколько по сфере приложения. Первая группа - статистический анализ *экспертных оценок*[12]. Многие изложенные выше методы могут использоваться для анализа экспертных оценок, а методы, развитые для анализа экспертных оценок, - к анализу других видов социологической информации. Тем не менее целесообразно выделить эту группу методов, так как специфика экспертных опросов (небольшие выборки при больших объемах информации, полученных от экспертов, и сложных процедурах опроса, наличие специальных процедур согласования мнения и т.п.) все же приводят к определенной специализации методов. С точки зрения математического обоснования экспертных оценок и обработки результатов экспертизы выделяют следующие виды задач[13] 1) построение моделей, описывающих поведение эксперта (модель поведения экспертов при ранжировании объектов,

[218]

[12] Статистические методы анализа экспертных оценок. М., 1977.

[13] Шмерлинг Д. С. и др. Экспертные оценки. Методы и применение (обзор). - В кн.: Статистические методы анализа экспертных оценок. М., 1977, с. 307.

модель парных или множественных сравнений[14] и т.п.); 2) проверка адекватности модели данным, полученным от экспертов; 3) оценка степени согласованности мнений экспертов; 4) получение коллективного мнения группы экспертов.

Вторая группа методов – анализ социометрических данных[15]. Под социометрическими обычно понимают методы исследования структуры межличностных отношений в малой социальной группе путем изучения выборов, сделанных членами группы по тому или иному критерию. Социометрические данные можно представить в виде графа, вершинами которого являются члены группы, а дугами - сделанные ими выборы. Более широко социометрические методы можно определить как методы сбора и анализа информации, представленной в виде графа, вершинами которого являются члены группы. Специфика здесь связана с тем, что результатом социометрического опроса является не значение признака, а выявление взаимоотношений индивидов между собой. При этом возникают задачи описания характеристик вершин графа (т.е. расчет так называемых индивидуальных социометрических индексов), характеристик структуры графа (групповые социометрические индексы и методы выделения подструктур – лидера, подгруппы, ослабляющих и укрепляющих членов группы и т.п.), описания связи между графами, построенными по разным критериям. Весьма специфичны методы поверки гипотез, основанные на проверке соответствия случайных графов с графами, полученными в исследовании.

Наконец, статистические методы используются также при моделировании социальных объектов, но рассмотрение этого вопроса выходит за рамки книги.

[219]

[14] Дзвид Г. Метод парных сравнений. М., 1978; Паниотто В. И. Метод множественных сравнений. - Социологические исследования, 1980, №1.

[15] Паниотто В. И. Структура межличностных отношений. Методика и математические методы исследования. Киев, 1975.

Глава VII

ИСПОЛЬЗОВАНИЕ ПРОГРАММИРУЕМЫХ МИКРОКАЛЬКУЛЯТОРОВ ДЛЯ АНАЛИЗА СОЦИОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

1. Организация обработки социологической информации. Классы задач, решаемых на ЭВМ и на программируемых микрокалькуляторах

Бурное развитие микроэлектроники в последние годы привело к возникновению нового класса клавишных вычислительных машин – программируемых микрокалькуляторов, которые практически воплотили «...еще недавно казавшуюся фантастической мечту об ЭВМ в кармане»[1]. Наряду с карманными появляются также новые типы программируемых настольных вычислительных машин, бурно развивается производство микро- и мини-ЭВМ, отличающихся от больших ЭВМ меньшими объемом памяти и скоростью выполнения операций, но значительно более дешевыми и надежными.

Анализ тенденций развития программируемых клавишных вычислительных машин показывает, что они являются перспективной группой средств вычислительной техники. Увеличение их вычислительной мощности привело к стиранию граней между ними и мини-ЭВМ[2], в свою очередь приближающихся по возможностям к большим ЭВМ[3]. Таким образом, различные типы вычислительной техники начинают равномерно заполнять разрыв между обычными калькуляторами и большими ЭВМ.

Все это приводит к изменению стратегии использования вычислительной техники. Если раньше основным был принцип централизованной обработки информации, то теперь возникает новый принцип распределенной (рассредоточенной)

[220]

[1] Трохименко Я. К., Любич Ф. Д. Инженерные расчеты на микрокалькуляторах. Киев, 1980, с. 6.

[2] Ландеховская Н. Г., Меньшикова Л. А. Современное состояние и тенденции развития программируемых ЭВМ. Информационный бюллетень «Приборы, средства автоматизации и системы управления». Серия ТС-2. Средства вычислительной техники и оргтехники. М., 1979.

[3] Брусенков Н. П. Миникомпьютеры. М., 1979, с. 20.

обработки. «Согласно этому принципу процессорная мощность цифровой системы не концентрируется в одном месте, а рассредоточивается так, чтобы была вблизи ее *потребителей* (курсив наш. - Авт.). Данное превращение структуры системы напоминает изменение характера электропривода, произошедшее в 30-х годах, когда один большой электромотор с ременными передачами к станкам был заменен индивидуальными для каждого станка электродвигателями»[4].

Социологическая информация не является исключением и перспективы совершенствования методов ее анализа связаны не столько с разработкой пакетов прикладных программ для больших ЭВМ с возможностью диалоговой работы, как полагают некоторые авторы[5], сколько с совершенствованием организационных и технических средств сочетания обработки на больших ЭВМ и на микро-, мини-ЭВМ и программируемых клавишных ЭВМ (ПЭКВМ). Наш опыт организации обработки социологической информации показал, что для решения широкого класса задач даже относительно ограниченные по возможностям программируемые микрокалькуляторы (типа «Электроника БЗ-21» и «Электроника БЗ-34») значительно эффективнее, чем ЭВМ. Их использование позволяет внести существенные изменения в организацию обработки социологической информации.

С определенной долей условности обработку информации можно разделить на первичную и вторичную. Под первичной понимается обработка, исходной информацией для которой служат ответы респондентов (заполненные анкеты), первичная обработка представляет собой различного рода преобразования социологической информации: расчет одномерных и многомерных распределений признаков, таксономия, классификация и т.п. Результатом вторичной обработки являются показатели, рассчитанные на основе данных первичной или выполненной ранее вторичной обработки[6], т.е. показатели, рассчитываемые по частотам, сгруппированным данным и т.п. (средние, меры рассеивания, связи, показатели значимости).

[221]

[4] Брусенцов Н. П., Миникомпьютеры. М., 1979, с. 13.
[5] SPSS (Statistical package for the social sciences). McGrow - Hill, 1975, p. XXII.
[6] Обработка результатов вторичной обработки тоже может считаться вторичной (нет смысла вводить понятия третичной, четвертичной и прочих видов обработки).

Естественно, что вся первичная обработка производится на ЭВМ. Исключение могут составлять пилотажные исследования и экспертные опросы в тех случаях, когда число опрашиваемых не превышает 20 - 30 человек - в этом случае первичная обработка может производиться вручную. Что же касается вторичной обработки, то в настоящее время она тоже производится на ЭВМ, а класс задач, решаемых на калькуляторах, чрезвычайно узок (это преимущественно суммирование и расчет процентов). Между тем, как будет показано ниже, решение значительного класса задач вторичной обработки информации на калькуляторах намного более эффективно, чем на ЭВМ. Рассмотрим этот вопрос подробнее.

Вторичная обработка социологической информации чаще всего включает в себя расчет мер центральной тенденции, вариации и связи, расчет уровней значимости и некоторых специальных показателей[7]: различные индексы, например привлекательности профессий, удовлетворенности работой; расстояния между двумя рядами распределений и т.п. К вторичной обработке можно отнести также некоторые из методов, рассмотренных в предыдущей главе: факторный анализ, исходной информацией для которого выступает матрица корреляций, причинный анализ, некоторые из социометрических методов. Вопрос о том, какие из видов вторичной обработки целесообразно проводить на ЭВМ, а какие - на микрокалькуляторах, зависит, на наш взгляд, от объема исходной информации, по которой рассчитывается показатель (например, от размерности матрицы), от числа показателей, которые требуется рассчитать и от организации работы на ЭВМ в данном социологическом подразделении (свой ВЦ или арендуемый, есть ли возможность работы в диалоговом режиме и т.д.). Кроме того, если необходимо вычислить нестандартные, редко используемые показатели, для расчета которых исследователь не располагает программами для ЭВМ, - более целесообразно рассчитать их на микрокалькуляторе.

Дело в том, что вторичная обработка социологической информации - это итеративный процесс, тесно сливающийся с анализом информации. Ее можно описать следующей цепочкой: интерпретация данных первичной обработки – расчет показателей для проверки гипотез, возникших при этом, - интерпретация полученных показателей и выдви-

[222]

[7] Имеются в виду показатели специфичные именно для рассматриваемых социологами проблем и не общепринятые в статистике.

жение новых гипотез – расчет новых показателей и т.п. Для расчета показателей на ЭВМ могут понадобиться следующие виды работ: перенос необходимых данных на специальные бланки, перфорация и контроль перфорации, организация доступа к ЭВМ (от вызова необходимых программ и информации из банка данных при работе с диалоговым монитором до заказывания машинного времени), счет. В общей сложности от возникшей необходимости рассчитать некоторый показатель до его расчета может пройти от нескольких часов до нескольких дней. Исключение составляет, пожалуй, лишь работа с диалоговым монитором при наличии необходимых программ и свободного доступа к ЭВМ.

На микрокалькуляторе требуется лишь ввести программу (несколько минут) и вводить данные (с визуальным контролем по индикатору) непосредственно с клавиатуры в регистры памяти или в операционные регистры. Расчет одного показателя занимает от нескольких секунд до нескольких минут, поэтому в случае необходимости рассчитать небольшое число показателей (несколько десятков) ПКЭВМ значительно эффективней[8].

Разумеется, при необходимости рассчитать большое число показателей (например, матриц коэффициентов корреляции, содержащих сотни коэффициентов) следует использовать ЭВМ. Другой случай предпочтительного использования ПКЭВМ – расчет редко используемых показателей. Дело в том, что процесс программирования и отладки про, грамм на ПКЭВМ значительно проще, чем на обычных ЭВМ, поэтому в данном случае может иметь смысл обработка на ПКЭВМ и достаточно больших массивов информации. Другими словами, при оценке целесообразности выбора того суммарное время, затрачиваемое на создание программы, подготовку информации и другие этапы обработки.

Все эти соображения приводят к следующему разделению функций. На ЭВМ целесообразно рассчитывать статистику, сопровождающую таблицы одномерных, двухмерных и многомерных распределений признаков (меры центральной тенденции и меры вариации, рассчитываемые для каждой строки или каждого столбца таблиц, а также всевозможные

[223]

[8] Под эффективностью мы имеем в виду, прежде всего, экономию временных затрат; сказанное тем более касается финансовых затрат, так как стоимость, например, микрокалькулятора «Электроника БЗ-21» (80 р.) приблизительно равна стоимости аренды одного часа работы ЭВМ ЕС-1022.

коэффициенты корреляции для таблиц). Кроме того, на ЭВМ целесообразно рассчитывать матрицы коэффициентов связей и уровней значимости признаков, например, матрицы коэффициентов связи между строками таблиц двухмерных распределений признаков. На микрокалькуляторах предпочтительно рассчитывать все показатели, определяемые не по таблицам сопряженности, а по отдельно взятым признакам или парам признаков, в частности коэффициенты корреляции для двух или нескольких признаков, меры значимости корреляций и различий между показателями. Кроме того, на калькуляторах целесообразно рассчитывать меры центральной тенденции, вариации и различные индексы, исходной информацией для которых служат средние коэффициенты корреляции и другие вторичные показатели, сопровождающие таблицы распределений признаков.

Это разделение (как и приводимые в следующем параграфе программы) относятся к использованию самого распространенного и доступного типа ПКЭВМ - «Электроники Б3-21», поступающего в свободную продажу в магазины канцтоваров. Использование ПКЭВМ, обладающих более широкими возможностями, разумеется, расширяет класс задач, решаемых на программируемых клавишных ЭВМ. Так, например, использование настольной ПКЭВМ типа «Искра-125» (ввод программы с длиной до 100 шагов с магнитных карт, а исходных данных – с накопителей на магнитных лентах) или карманного микрокалькулятора фирмы Hewlett Packard типа HP-41C (программы длиной до 2000 шагов вводятся с магнитных карт, память для данных -319 ячеек, наличие печатающего устройства[9]) позволяет обрабатывать по относительно простым программам большие массивы информации (например, рассчитывать матрицы коэффициентов корреляции, проводить факторный анализ). Более совершенные ПКЭВМ (например, настольная клавишная ЭВМ типа HP-9830B с оперативной памятью, приблизительно равной памяти «Минск-22», устройствами записи и считывания с магнитных лент, перфокарт, с печатающим устройством, дисплеем и графопостроителем[10]) позволяют выполнять все виды вторичной и многие из видов первичной обработки информации.

Можно предположить, что более перспективной является такая организация обработки информации, при которой в

[224]

[9] Hewlett Packard. Electronic instruments and systems, 1980, # 4, 1980. (Каталог продукции).
[10] Там же.

социологических подразделениях располагается так называемая станция клавишного ввода[11], представляющая собой ПКЭВМ или микро-ЭВМ, снабженную дисплеем и устройством записи информации на магнитную ленту. Информацию прямо с анкет (минуя кодирование и перфорацию) вводят в ПКЭВМ или микро-ЭВМ, проверяют, редактируют и записывают на магнитную ленту (или магнитный диск). Эта лента переносится затем на большую ЭВМ, на которой производится первичная и часть видов вторичной обработки информации, результаты которой частью печатаются, а частью записываются на магнитную ленту. Затем лента опять переносится на ПЭКВМ, на которой производится детальный и углубленный анализ полученной информации, проверка гипотез, возникших при первичном анализе информации, расчет новых показателей и т.п.

2. Программы расчета статистических мер и уровней значимости

Изложенные ниже программы предназначены для работы на программируемом микрокалькуляторе «Электроника-БЗ-21» (и могут с незначительными изменениями использоваться для работы на ПКЭВМ «Электроника БЗ-34»). Микрокалькулятор «Электроника БЗ-21» функционирует в двух режимах. Нажатием клавиш Р и РП он переводится в режим программирования (рис. 30), во время которого в калькулятор вводится программа (максимальная длина программы – 60 шагов, имеются команды условного и безусловного перехода и возможность использовать подпрограммы, а также 7 ячеек обычной и 6 – так называемой «стековой» памяти). Затем нажатием клавиш Р и РР калькулятор переводится в режим работы, во время которого он автоматически производит расчеты по введенной программе или используется в качестве обычного микрокалькулятора.

Программы 2, 5, б, 7 написаны Г. П. Талантом, программы 8, 9 и частично 1 и 11, а также использованные нами обозначения, заимствованы у Л. И. Францевича[12]. Отметим,

[225]

[11] Брусенцов Н. П. Миникомпьютеры, с. 28.

[12] Францевич Л. П. Обработка результатов биологического эксперимента на микро-ЭВМ «Электроника БЗ-21». Киев, 1979. Эту же книгу можно рекомендовать желающим освоить программирование на этом микрокалькуляторе. Наш опыт ведения семинаров для программирования на «Электронике БЗ-21» показывает. что для обучения программированию достаточно 4.х занятий.

что для работы по приведенным ниже программам желательно знакомство с инструкцией к микрокалькулятору, в частности, отметим, что появляющиеся на индикаторе микрокалькулятора числа иногда представлены в виде мантиссы и порядка числа. Например, запись [1.234567 03] на индикаторе означает 1,234567 * 10^3, или 1234,567, а запись [3,361255—02] означает 3,361255 * 10^{-2}, или 0,0361255. После ввода программы необходимо сначала провести расчеты для приведенного к каждой программе контрольного примера. Если полученный результат не совпадает с указанным в контрольном примере, то это означает, что при вводе

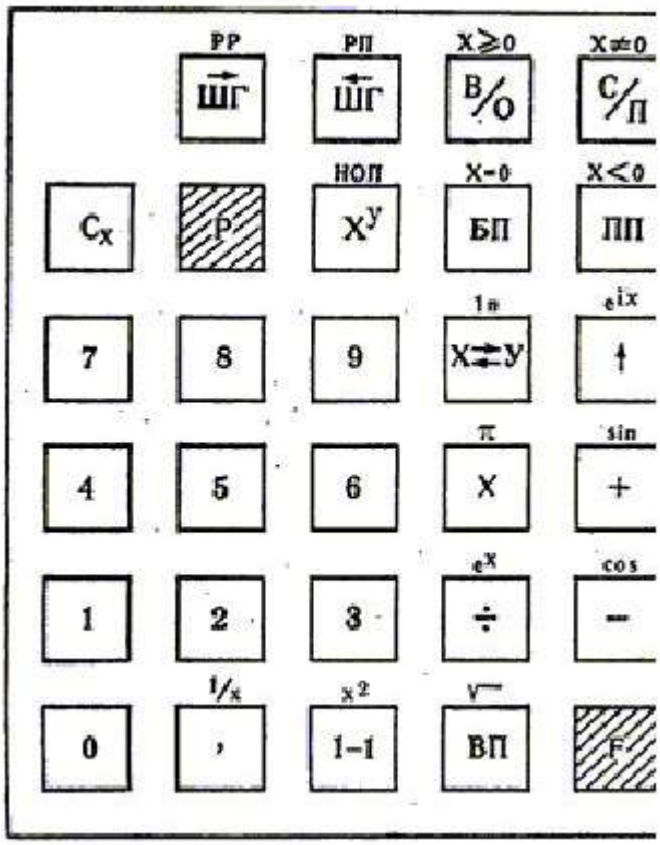


Рис. 30. Клавиатура микрокалькулятора «Электроника БЗ-21»

[226]

программы допущены ошибки, и программу следует ввести заново.

Необходимость создания программ специально для обработки данных социологических исследований связана с некоторыми особенностями вторичной обработки: большое количество порядковых и номинальных шкал, требующих использования ранговых и непараметрических критериев, наличие количественных признаков с заранее задаваемыми интервалами и т.д.

Другой особенностью уже технического, а не содержательного характера является своеобразие форм исходной информации – в таблицах одномерных, двухмерных и трехмерных распределений, использующихся в социологических исследованиях, есть ряд уже вычисленных данных (проценты, суммы, коэффициенты), что дает возможность создавать более компактные программы и совмещать в одной программе расчет нескольких показателей.

При описании программ приняты следующие обозначения:

1) программы набираются по столбцам; после набора первого столбца на индикаторе справа появится число 10, после набора 2-го - 20, 3-го - 30 и т.д., что позволяет контролировать число введенных команд;

2) в прямоугольную рамку заключены операции, которые повторяются при вводе каждого числа или пары чисел из ряда исходных данных, пока ряд не исчерпается;

3) в круглых скобках помещены пояснения, в частности, описание выводимых на индикатор результатов и указание примерного времени автоматического счета;

4) обмен содержимым регистров *x* и *y* обозначен *xy*, а стековые операции обозначены с использованием соответствующих подписей на клавишах, т.е. Р,; Р /—/

5) запятая в текстах программ и инструкций для вычисления - это не разделитель двух команд, а название соответствующей клавиши (будьте внимательны, не пропускайте ее при наборе программ!).

Среднее арифметическое, дисперсия, среднеквадратическое отклонение, коэффициент вариации, оценка дисперсии

А. Несгруппированные данные.

Представление данных

Результаты измерения: $x_1, x_2, ..., x_i, ..., x_n$

[227]

Результаты

Количество наблюдений n, среднее арифметическое $\bar{x}$, дисперсия σ^2, среднеквадратическое отклонение σ, коэффициент вариации C_U оценка дисперсии s^2.

Ввод программы

В/О	C_x	Р /—/	+	БП	Р4	F2	+	F√	F5	Р
Р	Р2	+	Р2	F↑	С/П	+	÷	С/П	С/П	РР
РП	Р3	Р,	F3	xy	Fx²	↑	↑	↑		
	Р,	×	1	Р /—/	×	Р,	F5	F4		
	С/П	↑	+	÷	/—/	÷	×	÷		
	↑	F2	Р3	$F^1/_x$	↑	Р5	С/П	С/П		

Вычисление

Нажимаем клавиши В/О С/П [x_i С/П] (значение *n,* время счета 2 - 3 сек). Напоминаем, что запись [x_i С/П] означает x_1 С/П x_2 С/П..., x_n С/П. Значение *i*, появляющееся на индикаторе после ввода каждого числа, помогает оператору контролировать, сколько чисел он уже ввел. После ввода всех чисел вычисляются показатели: БП Р ÷ С/П (значение $\bar{x}$), С/П (значение σ^2), С/П (значение σ), С/П (значение C_U), С/П (значение s^2).

Контрольный пример

Исходные данные: x_i = 11 12 13 14 15

Результаты: $n = 5$, $\bar{x} = 13$, σ^2=2, σ = 1,414213, C_U = 0,1087856, s^2 = 2,5

Формулы расчета

См.: (формулу с. 39), (I,4,1), (с. 40), (I,4,3).

Б. Сгруппированные данные.

Представление данных

Границы класса $...x' - x''...$,(или среднее значение $...x_i...$), количество наблюдений $...N_i...$

Результаты

Количество наблюдений n, среднее арифметическое $\bar{x}$, дисперсия σ^2, среднеквадратическое отклонение σ, коэффициент вариации C_U, оценка дисперсии s^2.

Ввод программы

В/0	C_x	F8	↑	↑	F4	P / — /	×	P,	F5	F4	P
P	P2	÷	P / — /	F2	+	÷	/ — /	÷	×	÷	PP
РП	P3	↑	+	+	P3	$F^1/_x$	↑	P5	С/П	С/П	
	P,	P7	P,	P2	БП	P4	F2	+	F√	F5	
	С/П	F4	F7	F3	F ↑	С/П	+	÷	С/П	С/П	
	P4	×	×	↑	—	Fx^2	↑	↑	↑		

[228]

Ввод константы

Если задано среднее значение класса, то 1Р8;

если заданы границы, то *2Р8.*

Вычисление

В/О С/П (не обращайте внимания на числа, которые могут появиться на индикаторе). Если задано среднее значение класса: $x_i' \uparrow N_i$ С/П. Если заданы границы класса: $x_i' \uparrow x_i'' + N_i$ С/П (получим значение n).

1 БП, (запятая означает, что надо нажать клавишу, на которой изображена запятая) С/П (значение $\bar{x}$) С/П (значение σ^2) С/П (значение σ) С/П (значение C_U) С/П (значение s^2)

Контрольный пример

Исходные данные

x_i	11	12	13	14	15
N_i	1	3	10	3	1

Ввод констант: 1 Р8 $\sigma^2 = 0.7777777$ $\sigma = 0.8819171$

Результаты: $n = 18; \bar{x} = 13; C_U = 0.06783977; s^2 = 0.8235294$

Формулы расчета: См.: (с. 39), (I,4,3)

Квантили, медиана для сгруппированных данных (при равных интервалах)

Представление данных

Границы класса $(x_i' - x_i'')$, расположенные в порядке убывания, N_i - количество наблюдений (частоты), N - число опрошенных (сумма частот).

Результаты

Квантиль P_p (p - доля, для которой нужно определить квантиль), медиана (рассчитывается как p = 0,50).

Ввод программы

В/О	↑	С/П	Px < 0	+	—	×	БП	P
P	F7	P5	F÷	P,	xy	↑	F / — /	PP
PП	×	+	xy	БП	F5	P / — /		
	P6	↑	P / — /	P1	÷	—		
	C_x	F6	↑	xy	↑	С/П		
	↑	—	F8	F5	F8	xy		

Вычисление

1. Занести в регистр «Р8» ширину интервала: $x_1' \uparrow x_1''$ — Р8 (в случае, если таблица составлена так, что $x_1'' \neq x_{i+1}$, например, 20 — 23, 24 — 27, 28 — 31 и т.д., прибавляется единица, т.е. $x_1'' \uparrow x_1' - 1 +$ Р8).

[229]

2. $x_1' P$ (ввод нижней границы в стековый регистр[13]) NP7

3. Набираем величину квантиля: p В/О С/П (на индикаторе 0; 2 сек)

4. [N_t С\П] (на индикаторе будут появляться накопленные частоты; 2 сек.). В процессе вычисления калькулятор выделяет из введенных частот ту, которая соответствует интервалу, содержащему квантиль, и автоматически переходит к его вычислению. При этом (чтобы отличить его от накопленных частот) квантиль появляется со знаком «—». Таким образом, первое отрицательное число, появившееся на индикаторе, представляет собой искомое

[13] В случае, если $x_i'' \neq x_{i-1}'$, x_i' заменяем на $x_i' - \frac{1}{2}(x_2' - x_1'')$, т.е. вводим C_U.

значение квантиля *Pp*. Никакого другого смысла минус не имеет и его следует отбросить. В случае, если оператор не обратит внимания на то, что на индикаторе появилось отрицательное число и будет продолжать вводить N_t С\П на индикаторе все равно будет восстанавливаться значение полученного квантиля.

5. Если по той же таблице необходимо рассчитать несколько квантилей, то можно поступить двояким способом. Во-первых, можно повторять пп. 3 и 4 заново. Во-вторых, можно начать вычисления с наибольшего квантиля, записывая накопленные частоты. В этом случае вычисления последующих квантилей производится следующим образом: *F6* (на индикаторе появляется значение произведения *pN).* По этой величине определяем из таблицы соответствующие N_i. и нижнюю границу N_t. Затем $N_i P5 x_i' P$ (не обращать внимания на появляющиеся при этом числа). Определяем по таблице предыдущую накопленную частоту $cumF_{i-1}$. Далее $cumF_{i-1}\uparrow$ *F6* БП Р4 С/П (получим значение квантиля; 2 сек).

Контрольный пример

Границы интервалов	$x_i' - x_i''$	20-24	24-28	28-32	32-36	36-40	…	N
Частоты	σ	135	295	291	307	204	…	204

Медиана $P_{0.50} = 35.51791$ («—» отбрасывается).

[230]

Формулы расчета[14]

$$Pp = x_i + \frac{pN - cumF_{i-1}}{N_i}\left(x_1'' - x_1'\right),$$

где x_i' - нижняя граница интервала, содержащего, частоту pN

N_i - частота этого интервала

$cumF_{i-1}$ - накопленная к *i*-му интервалу.

Энтропийная мера вариации

Представление данных

$N_1, N_2, ..., N_k$ - частоты одномерного распределения признака

Результаты

Энтропийная мера вариации c_v, сумма частот N

Ввод программы

В/О	Cx	↑	P3	F5	F3	P_{ln}	/—/	P
P	P3	P_{ln}	F2	1	↑	—	С/П	PP
РП	P4	×	↑	+	F4	↑		
	P5	↑	F4	P5	+	F5		
	С/П	F3	+	БП	↑	P_{ln}		
	P2	+	P4	F ↑	F4	+		

Вычисление

В/О С/П 2 сек. [N_i С/П (5 сек., на индикаторе значение i)] БП Р4 С/П (получим $\cdot N$; 5 сек.) F4 (получим N).

Контрольный пример

Варианты		Результат y_2
Частоты	5 7	*N*=39

Формулы расчета

$$N = \sum_{i=1}^{k} N_i$$

$$\varepsilon = -\frac{1}{\ln k}\sum_{i=1}^{k}\frac{N_i}{N}\ln\frac{N_i}{N} = -\frac{1}{\ln k}\left(\frac{1}{N}\sum_{i=1}^{k} N_i \ln N_i - \ln N\right)$$

[231]

[14] В случае, если $x_i'' \neq x_{i+1}'$, формула несколько видоизменяется: x заменяется на $x_i'' \neq x_{i-1}'$, а $\left(x_1'' - x_1'\right)$ на $\left(x_1'' - x_1' + 1\right)$: программа остается без изменений, требуется лишь выполнить указания в примечаниях к пп. 1 и 2.

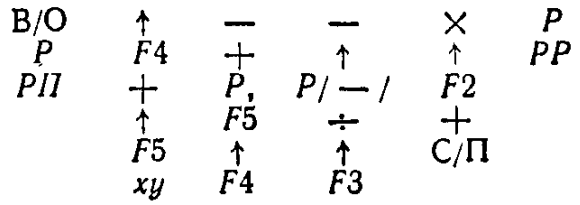
Мода для сгруппированных данных

Представление данных

Дано одномерное распределение признака X, N_i - частоты, x_i' и x_i'' - границы класса, I_l - ширина интервала(см. сноску к предыдущей программе), l - интервал, содержащий моду.

Результаты: Мода (Мо)

Ввод программы



Вычисление

$x_l P2 I_l P3 N_{l-1} P4 N_l P5 N_{l+1} B/0 C/П(M0;4сек.)$

Контрольный пример

Признак	60-80	80-100	100-120	120-140
Частоты		110	204	50

Мо=107,5806

Формулы расчета – см.: (I,3,3).

Коэффициент корреляции r, уровень значимости t

Представление данных

Число объектов n, $x_1, x_2, ..., x_n$ - значение переменной X, $y_1, y_2, ..., y_n$ значения переменной Y для каждого объекта, N_x и N_y - сумма значений переменной X и Y соответственно.

$$N_x = \sum_{i=1}^{n} x_i \quad N_y = \sum_{i=1}^{n} y_i$$

Результаты

Коэффициент корреляции между признаками X и Y – r, уровень значимости t.

[232]

Ввод программы

В/О	*P3*	*F¹/x*	×	С/П	*F5*	*F6*	*F7*	↑	*xy*	*F4*	*P*
P	*xy*	*xy*	*P6*	*P3*	+	+	+	*F7*	÷	2	*PP*
PП	*P2*	×	*F3*	*xy*	*P5*	*P6*	*P7*	×	С/П	+	
	×	*P5*	*Fx²*	*P2*	*F2*	*F3*	БП	*F√*	1	*F¹/x*	
	↑	*F2*	×	×	*Fx²*	*Fx²*	*P3*	↑	—	БП	
	F4	*Fx²*	*P7*	↑	↑	↑	*F6*	*F5*	*P7*	*P7*	

Вычисление

n | — | P4 N_x ↑ N_y В/О С/П (4 сек.)

$\left[x_i \uparrow y_i C/П\right]$ (4 сек). Не обращайте внимания на появляющиеся при этом на индикаторе числа. После ввода всех пар значений переменной *X* и *Y* определяем *r* и *t*. БП ВП С/П (получаем r; 3 сек). P5 Fx^2 С/П (получаем уровень значимости t; 4 сек).

Контрольный пример

Исходные данные:

$n=5$	x_i	11	12	13	14	15	$N_x = 65$
	y_i	22	22	23	25	25	$N_y = 117$

Результаты

$r = 0.9383148 = 9.383148 \cdot 10^{-1}$,

$t = 4.700095$

Так как при 3-х степенях свободы критическое значение *t* на уровне 5% равно 3,2, а на уровне 1% - 5,8, данное значение r значимо на уровне 5%.

Формулы расчета См. (II,5,3), (V,8,1)

Критические значения

См. табл. И Приложение 3.

Коэффициент корреляции Спирмена (ρ) *при отсутствии связанных рангов*[15]*, уровень значимости t*

Представление данных

n - число объектов,

$R_1^{(x)}, R_2^{(x)}, ..., R_n^{(x)}$ - ранги 1-го, 2-го, …, n-го объекта соответственно по переменной *X*

$R_1^{(y)}, R_2^{(y)}, ..., R_n^{(y)}$ - ранги 1-го, 2-го, …, n-го объекта соответственно по переменной *Y*

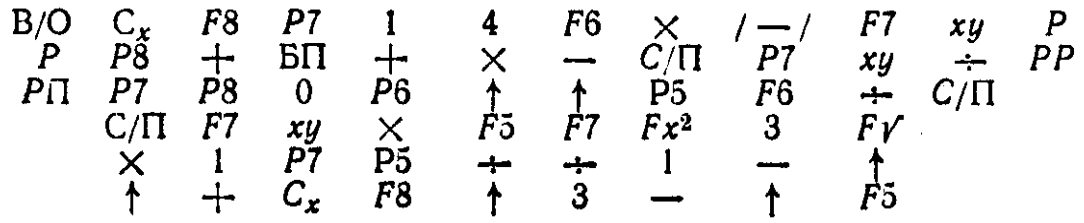
[233]

[15] В случае связанных рангов следует воспользоваться программой для вычисления *r* - это всегда позволит определить ρ независимо от того, связаны ранги или нет (гл. 11, §4).

Результаты

Количество введенных пар рангов n (для контроля правильности ввода), коэффициент корреляции ρ, уровень значимости t

Ввод программы.



Чистка

В/О С/П

Вычисление

[$R_i^{(x)} \uparrow R^{(y)}{}_i$ С/П] (после ввода каждой i-й пары на индикаторе появляется текущее значение i, что облегчает оператору правильность ввода; 3 сек). После ввода всех пар чисел на индикаторе должно появиться заданное число n.

БП 2 С/П (получим ρ; 5 сек),

С/П (получим t; 5 сек).

Контрольный пример

Исходные данные:

n=12

$R_i^{(x)}$	2	8	12	3	1	6	7	10	4	9	11	5
$R_i^{(y)}$	6	5	10	7	3	4	9	8	1	11	12	2

Результаты

$\rho = 0.7062935$

$t = 3.155016$

Формулы расчета

См. (V,8,4)

Критические значения. Табл. И, Приложение 3.

Коэффициент корреляции τ, уровень значимости z

Представление данных

n - число объектов

$x_1, x_2, ..., x_n$ - значения переменной X (или ранги по X) для каждого объекта

$y_1, y_2, ..., y_n$ - значения переменной Y (или же ранги по Y) для каждого объекта.

[234]

Результаты

Коэффициент корреляции τ, для несвязных и для связных рангов, уровень значимости **z** (для случая несвязных рангов).

Ввод программы

В/О	С/П	F3	F8	—	БП	÷	ПП	9	/—/	÷	P
P	P,	—	БП	×	2	↑	P8	÷	Px≠0	↑	PP
PП	F2	↑	PO	↑	P/—/	F7	С/П	↑	+	F7	
	—	P,	С/П	P/—/	×	xy	1	F8	↑	+	
	P/—/	×	↑	+	FV	÷	0	БП	×	P7	
	↑	ПП	1	P,	2	С/П	+	4	FV	В/О	

Вычисление

1. C_x P7 (чистка) БП FO x_1 P2 y_1 P3 [$x_i \uparrow y_i$ С/П] (4 сек.)

Действия рамке выполняются для i = 2, 3,…, n. Затем x_2 P2 y_2 P3 [$x_i \uparrow y_i$ С/П] (i = 3, 4,…n) x_3 P2 y_3 P3 [χ^2 С/П] (i = 4, 5,…n) и так далее.

Последний раз x_{n-1} P2 y_{n-1} P3 $x_n \uparrow y_n$ С/П

2. nP3 БП F ×, С/П (2 сек.) xy P8 /—/ P,

3. Пункт 3 выполняется только при наличии связанных рангов по X – в противном случае переходим сразу к п.6

4. F8 /—/ P

5. Пункт 5 выполняется лишь в том случае если есть связанные ранги по *Y*, в противном случае переходим сразу к п. 6

[g_i С/П] (2 сек.), g_i - число связанных рангов в i-группе связанных рангов по признаку *Y*

6. P /—/ ↑ БП F4С/П (значение τ; 3 сек.)

7. С/П (2 сек.) F3 4^y С/П (значение **z**; 4 сек.)

Контрольный пример

Исходные данные

x_i	5	1	3	4	2
y_i	3	1	4,5	4,5	2

n=5 *g*=2 для переменной *Y*

Результаты: $\tau = 0.5270463$, $z = 0.9797959$

Формулы для расчета[16]

$$\tau = \frac{P-Q}{\sqrt{\frac{n(n-1)}{2}-T_x}\sqrt{\frac{n(n-1)}{2}-T_y}} = \frac{S}{\frac{1}{2}\sqrt{\left[n(n-1)-\sum_i f_i(f_i-1)\right]\left[n(n-1)-\sum_i g_i(g_i-1)\right]}}$$

$$Z = \frac{S^*}{\frac{1}{18}n(n-1)(2n+5)} = \frac{S^*}{\frac{1}{2}\sqrt{\frac{4n+10}{9}n(n-1)}},$$

где $S^* \begin{cases} S+1 & при \quad S>0 \\ S & при \quad S=0 \\ S-1 & при \quad S<0 \end{cases}$

$S = P - Q$

Критические значения

[16] Формулы для **z** в случае связанных рангов см. в кн.: *Кендэл М.* Ранговые корреляции. М., 1975, с. 66.

См. табл. А Приложение 3.

Примечание Пункт 1 программы вычисления τ довольно трудоемкий, иногда имеет смысл определить значение S, вычисляемое в пункте 1, вручную (в этом случае объекты ранжируются по значениям одной из переменных). Далее S заносится в «Р7» и вычисление повторяется, начиная с п. 2.

Линейная корреляция и регрессия

Представление данных

$x_1, x_2, ..., x_n$ - значения признака X

$y_1, y_2, ..., y_n$ - значения признака Y

Средние x и y, коэффициент корреляции r, количество пар объектов n и параметры a и b линейного уравнения $y = ax + b$

Ввод программы

В/О	*С/П*	*P,*	*F2*	*P,*	*Fx²*	*P2*	P/—/	P7	↑	*xy*	*P*
P	*P2*	ПП	ПП	1	+	P/—/	ПП	*F3*	*F2*	P/—/	*PP*
PП	*xy*	*F4*	*F4*	+	*P,*	*P3*	C_x	ПП	÷	—	
	P3	*F2*	ПП	*P,*	↑	ПП	*F6*	*P8*	↑	/—/	
	×	ПП	P4	БП	*F3*	C_x	×	С/П	*F4*	↑	
	↑	*P4*	+	*PO*	В/О	*P6*	*FV*	*P4*	×	В/О	

Чистка

↑ /—/ (шесть раз)

Вычисление

В/О С/П [x_i ↑ y_i С/П] (6 сек1

Р /—/ (значение n) БП Р5 С/П (10 сек)

[236]

F7 ÷ (получим r) F6 ÷ (значение a) ↑ Р/—/ Р/—/ Р/—/ (значение $\bar{x}$) × ↑ Р/—/ Р/—/ (значение y) xy (значение b)

Контрольный пример

Исходные данные

x_i	11	12	13	14	5
y_i	22	22	23	25	25

Результаты:

$n = 5, r = 0.9383148, a = 0.9, \bar{x} = 13, y = 23.4, b = 11{,}7$

Формулы расчета (см. II,5,3), (III,1,11), (III,1,12).

Линейная корреляция и регрессия при группировке в классы

Представление данных

Частоты		N_1...	N_i...
Середины интервалов для признаков	X	x_1...	x_i...
	Y	y_1...	y_i...

Результаты

Общее число респондентов N, средние значение $\bar{x}$ и $\bar{y}$, коэффициент корреляции r и параметр линейного уравнения $y = ax + b$

Ввод программы

В/0	Р2	÷	÷	↑	↑	Р/—/	Р/—/	Р7	↑	*xy*	*P*
P	ПП	ПП	*F*4	ПП	*P*/—/	*F*2	ПП	*F*3	*F*2	*P*/—/	*PP*
PП	*P*4	*F*÷	ПП	*F*÷	+	В/0	C_x	ПП	÷	—	
	C/П	*C*/П	÷	С/П	↑	ПП	*F*6	*P*8	↑	/—/	
	*P*3	*P*4	*F*4	*F*3	*P*,	C_x	×	С/П	*F*4	↑	
	ПП	ПП	÷	×	*xy*	Р6	$F\sqrt{}$	*P*4	×	В/0	

Чистка

C_x Р /—/ (6 раз)

Вычисление [N_i В/О С/П (2-3 сек.) x_i С/П (4 сек.) y_i С/П (6 сек.)]

Р /—/ (получим N) Р2 Р/—/ Р3 БП 5С/П (10сек.)

F7 ÷ (получим r) F6 ÷ (получим a) ↑ Р/—/ Р/—/ Р/—/ (значение $\bar{x}$) × ↑ Р/—/ Р/—/ (значение y) xy (значение b).

[237]

Контрольный пример

N_i	10	3	3	10
x_i	10	10	20	20
y_i	10	20	10	20

Результаты:

$N = 26; r = 0.5384615; a = 0.5384615; \bar{x} = 15; \bar{y} = 15; b = 6.923077$

Формулы расчета

$N = \sum_{i=1}^{k} N_i; \sum_{i=1}^{N} x_i = \sum_{i=1}^{k} N_i x_i$ (k – число интервалов)

$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{k} N_i x_i^2; \sum_{i=1}^{N} y_i = \sum_{i=1}^{k} N_i y_i;$

$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{k} N_i y_i^2; \sum_{i=1}^{N} x_i y_i = \sum_{i=1}^{k} N_i x_i y_i;$

Остальные формулы те же, что и в предыдущем случае (т.е. без группировки в классы).

Программа для вычисления по таблице $k \times l$ значения χ^2, коэффициента Чупрова T, коэффициента Крамера T.

Исходные данные: Таблица $k \times l$ (см. § 2 гл. 11)

Результаты. Значение χ^2, T и T_c

Ввод программы

В/0	С/П	*F*2	БП	↑	↑	×	*P*6	$F\sqrt{}$	*P*
P	*xy*	÷	*P*0	*F*3	*F*5	↑	С/П	$F\sqrt{}$	*PP*
PП	Fx^2	↑	*F*8	×	×	*F*6	*F*4	↑	
	xy	*F*8	↑	*P*6	$F\sqrt{}$	*xy*	↑	*F*6	
	÷	+	1	*C*/П	↑	÷	*F*5	×	
	↑	*P*8	—	*F*4	*F*3	$F\sqrt{}$	÷	*C*/П	

Вычисление

Чистка C_x Р8. NP3. Большее из чисел $(k-1)$ и $(l-1)$ запоминаем в 4-й (Р4) меньшее - в 5-й ячейке (Р5).

Считаем по столбцам N (y_1) P2 N_{11} ↑ N (x_1) В/О С/П С/П и далее в цикле [N_{i1} ↑ $N(x_i)$С/П] i = 2, 3, ..., k (2 сек). Затем для второго столбца N (y_2) P2 [N_{i2} ↑ $N(x_i)$С/П] l = = 1, 2, ..., k и т.д. После ввода всех столбцов БП P × С/П (получим χ^2)

С/П (получим T) С/П (получим T^2).

[238]

Контрольный пример

	y_1	y_2	y_3	$N(x_i)$
x_1	29	36	15	80
x_2	14	24	2	40
$N(x_i)$	43	60	17	120

C_x P8 120 P3 2P4 1P5 Далее 43P2 29 ↑ 80 В/О С/П С/П 14 ↑40 С/П 60P2 36 ↑ 80 С/П 24 ↑ 40 С/П 17P2 15 ↑ 80 С/П 2 ↑ 40 С/П Все столбцы введены ВП Р × С/П (получим χ^2 = 4,770432) С/П (Т = 0,1676604) С/П (T_c = 0.1993830)

Формулы расчета
См. (II,2,1), (II,2,4), (II,2,5)
Уровни значимости
Табл. Б Приложение 3.

Коэффициенты связи для таблицы $2\times 2: Q, \Phi$ *и* χ^2

Представление данных:

	y_1	y_2
x_1	N_{11}	N_{12}
x_2	N_{21}	N_{22}

Результаты
Коэффициент Юла Q, коэффициент контингенции Ф, показатель χ^2.
Ввод программы

В/О	*xy*	↑	*F3*	*F6*	+	*F3*	↑	−	*xy*	÷	*P*
P	+	*F6*	+	×	↑	↑	*F5*	*P8*	÷	*С/П*	*PP*
РП	*P6*	×	↑	*P6*	*F6*	*F4*	×	−	*С/П*	↑	
	P7	*P6*	*F7*	*F2*	×	×	*P5*	+	*F8*	*F7*	
	F3	*F2*	+	↑	*F√*	*P4*	↑	↑	↑	×	
	+	↑	*P7*	*F4*	*P6*	*F2*	*F4*	*F8*	*F6*	*С/П*	

Вычисление $N_{11}P2$ $N_{12}P3N_{21}P4N_{22}P5$ В/О С/П (получили Q; 8 сек.) С/П (получили Ф; 2 сек.) Fx^2 С/П (получили χ^2; 1 сек.)
[239]

Контрольный пример:

Исходные данные: таблица

1	2
3	4

Результаты:

$Q = -0{,}20 \quad \Phi = -0.08908708 \quad \chi^2 = 0.07936508$

Формулы расчета

См (с. 86), (с. 88), (II,2,1).

Коэффициенты частной корреляции (для r и τ).

Представление данных:

Коэффициенты парной корреляции r_{01} , r_{02} и r_{12} (или τ_{01}, τ_{02} и τ_{12}) для коэффициента частной корреляции $r_{01.2}$ или $\tau_{01.2}$; коэффициенты $r_{01.2}, r_{03.2}$ и $r_{13.2}$ для коэффициента $r_{01.23}$ и т.д.

Результаты

Коэффициенты частной корреляции любого порядка (для вычисления коэффициентов n-ого порядка сначала вычисляются коэффициенты ($n-1$)-го порядка)

Ввод программы

В/О	*P*/—/	×	1	↑	×	С/П	*P*
P	↑	↑	↑	*F*3	*F*√		*PP*
PП	*Fx²*	*P*/—/	*F*2	*Fx²*	↑		
	*P*2	*xy*	—	—	*F*4		
	P/—/	—	*P*2	↑	*xy*		
	*P*3	*P*4	1	*F*2	÷		

Вычисление.

Для коэффициентов первого порядка $r_{01}P, r_{02}P, r_{12}P,$ В/О С/П (получим $r_{01.2}$, 7 сек).

Для коэффициентов второго порядка $r_{01.3}P, r_{02.3}P, r_{13.2}P,$ В/О С/П (получим $r_{01.23}$;7 сек) и т.д.

Аналогично вычисляются коэффициенты $\tau_{01.2}; \tau_{01.23}$ и т.д.

Контрольный пример

$r_{01.3} = 0.620 \quad r_{03.2} = 0.240 \quad r_{13.2} = -0.171 \quad r_{01.23} = 0.6911214$

Формулы расчета

(III,2,3 - III,2,5), (III,2,6), (III,2,7)

Коэффициент множественной корреляции

Представление данных

Частные коэффициенты корреляции, необходимые для расчета коэффициента множественной корреляции $R_{0.12\ldots n} : r_{01}; r_{02.1}; r_{03.12}; \ldots; r_{0n.12\ldots(n-1)}$

[240]

Результаты

Коэффициент $R_{0.12...n}$

Ввод программы.

В/О	1	xy	БП	F√	P
P	$P2$	—	$P\uparrow$	$C/П$	PP
$PП$	С/П	$\uparrow$	1		
	Fx^2	$F2$	$\uparrow$		
	$\uparrow$	×	$F2$		
	1	$P2$	—		

Вычисление

В/О С/П r_{01} С/П (3 сек) $r_{02.1}$ С/П (3 сек)... $r_{0n.12...(n-1)}$ С/П (3 сек) После ввода всех частных коэффициентов корреляции БП Р × С/П (получим $R_{01.2...n}$;3 сек)

Контрольный пример.

$r_{01}=0.705$ $r_{02.1}=0.479$ $r_{03.12}=0.342$ $R_{01.23}=0.8110240$

Формулы расчета

(III,3,7)

Значимость различия долей (процентов)

Представление данных

v_1, v_2 - сравниваемые доли признаков (проценты)

n_1, n_2 - объемы выборок, по которым рассчитывались v_1 и v_2 соответственно.

Результаты

Критерий значимости различий долей z (может использоваться, если выполняется каждое из следующих условий: $n_1 \geq 50, n_2 \geq 50, np > 5, n(1-p) > 5$, формула для вычисления приведена ниже).

Ввод программы

В/О	$P4$	$F1/x$	$F3$	$\uparrow$	$F7$	$/—/$	×	Fx^2	P
P	xy	$P7$	$F1/x$	$P/—/$	+	$\uparrow$	$F\sqrt{}$	$F\sqrt{}$	PP
$PП$	$P5$	$\uparrow$	$P4$	+	$\uparrow$	$F8$	$\uparrow$	$C/П$	
	—	$F4$	$\uparrow$	$P5$	$F5$	+	$F6$		
	$P6$	×	$F5$	$F4$	xy	$\uparrow$	xy		
	$F2$	$P,$	×	$\uparrow$	÷	$F5$	÷		

Счет

Если v_1 и v_2 выражены в долях, вводим: 1Р8, если в процентах: 100Р8

$n_1 P2$ $n_2 P3$ $v_1 \uparrow v_2$ В/О С/П (получим z; 9 сек)

Контрольный пример

Исходные данные:

[241]

$n_1 = 392$, $v_1 = 0.296$, $v_1 = 29.6\%$,
$n_2 = 277$ $v_2 = 0.209$ или $v_2 = 20.9\%$,

Результат: z = 2,526962
Формулы расчета
См (V,5,1).
Критические значения
См. табл. А Приложения 3.

Значимость различий средних арифметических

Представление данных
$\bar{x}_1, s_1, \bar{x}_2, s_2$ - средние арифметические и оценки средних квадратических отклонений, рассчитанные по выборкам объема n_1 и n_2 соответственно.
Результаты,
Критерий значимости различий t, число степеней свободы f
Ввод программы

В/0	C_x	С/П	F5	Fx^2	С/П	Fx^2	1	↑	P
P	P3	ПП	÷	F√	Fx^2	P4	+	F5	PP
РП	P5	F4	2	↑	xy	F3	↑	+	
	С/П	F3	—	F3	P2	+	F4	P5	
	ПП	Fx^2	С/П	F√	÷	P3	xy	В/0	
	F4	↑	—	÷	↑	F2	÷		

Вычисление
В/О *С* П (1 сек) $n_1 \uparrow s_1$ С/П ($\approx$4 сек)
$n_2 \uparrow s_2$ С/П (получим v; $\approx$5 сек)
$\bar{x}_1 \uparrow \bar{x}_2$ С/П (получим критерий t; 1 сек)
Контрольный пример
Исходные данные:
$\bar{x}_1 = 15, s_1 = 4, \bar{x}_2 = 11, s_2 = 0.1, n_1 = 30, n_2 = 65$
Результаты:
v = 29,01788
t = 5,476435
Формулы расчета
См. (V,6,1), (V,6,2)
Критические значения
См. табл. А Приложение И.

Значимость различий двух коэффициентов корреляции

Представление данных
Коэффициенты корреляции r_1 и r_2 рассчитанные по выборкам n_1 и n_2 соответственно. [242]

Результаты
Уровень значимости различий коэффициентов корреляции z.
Ввод программы

В/О	↑	3	*FV*	↑	*FV*	1	*P*
P	3	—	*P2*	*P,*	↑	+	*PP*
РП	—	*F1/x*	ПП	÷	*F2*	2	
	F1/x	↑	,	*FV*	÷	—	
	P2	*F2*	ПП	*Pln*	*C/П*	÷	
	P/—/	+	,	Fx^2	*P/—/*	В/О	

Вычисление
Не обращая внимания на числа, которые могут появиться из стека, выполняем:
r_1P, r_2P, n_1P, n_2 В/О С/П (получим *z*; 9 сек)
Контрольный пример
Исходные данные:

$r_1 = 0.6$ $\quad n_1 = 28$
$r_2 = 0.8$ $\quad n_2 = 23$

Результаты: z = 1,351550
Формулы расчета
См. (V,9,1).
Критические значения
См. табл. А приложения 3

Значимость различий распределений (критерий χ^2)

Исходные данные
Таблица распределений двух совокупностей респондентов по некоторому признаку X:

$k_1k_2...k_n$	K
$l_1l_2...l_n$	L

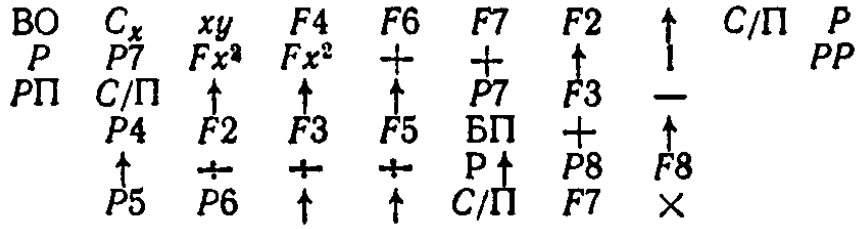
где k_i и l_i - частоты; k_i - число респондентов из первой совокупности, выбравших *i*-ю градацию при ответе на вопрос X; l_i - аналогичная частота для второй группы респондентов. *K* и *L* -- суммы частот:

$$K = \sum_{i=1}^{n} k_i \quad L = \sum_{i=1}^{n} l_i$$

Результаты
[243]

Критерий χ^2, показывающий значимость различий распределений признака X в 2 группах респондентов[17].

Ввод программы



Вычисление

Заносим K в Р2; L в Р3. В/О С/П (1 сек)

[$k_i \uparrow l_i$ С/П] ($t_1 \approx$ 4 сек) БП Р5 С/П (получим χ^2)

Контрольный пример

Исходные данные:

29	36	15	80
14	24	2	40

Результат: $\chi^2 = 4{,}770444$

Формулы расчета

$$\chi^2 = (K+L)\left[\sum_{i=1}^{n}\left(\frac{k_i^2}{K}+\frac{l_i^2}{L}\right)\frac{1}{k_i+l_i}-1\right]$$

Критические значения

См. табл. Б Приложение 3.

Что касается расчета специальных показателей, то во многих случаях (в частности, при расчете индексов удовлетворенности, престижа и т.п.) можно воспользоваться программой вычисления средних арифметических для сгруппированных данных (вместо середин интервалов взять значения весов тех или иных вариантов ответа). При большом числе индексов при неизменных весах для ускорения расчетов целесообразно использовать специальные программы. Приведем пример такого рода программы.

Расчет индекса для пятибалльных шкал

Представление данных

Одномерное распределение по признаку, имеющему 5 вариантов частоты: N_1, N_2, N_3, N_4, N_5. Сумма частот равна N. Каждый из вариантов имеет вес: $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$

[244]

[17] В цит. книге Л. Францевича есть программа расчета критерия в случае, если значение K и L не заданы.

Результат и формула расчета

Некоторый индекс

$$I = \frac{1}{N}\sum_{i=1}^{5}\beta_i N_i$$

Ввод программы

В/О	*Cx*	*F5*	*C/П*	↑	*F6*	*C/П*	В/О	Р
P	*P8*	*C/П*	↑	*F5*	ПП	×		*PP*
РП	*C/П*	↑	*F4*	ПП	*F5*	↑		
	↑	*F3*	ПП	*F5*	↑	*F8*		
	F2	ПП	*F5*	*C/П*	*F7*	+		
	ПП	*F5*	C/П	↑	÷	*P8*		

Вычисление

а) $\beta_1 P2$ $\beta_2 P3$, $\beta_3 P4$ $\beta_4 P5$ $\beta_5 P6$,

б) NP7

в) В/О С/П (1 сек.) N_1 С/П (3 сек) N_2 С/П (3 сек) N_3 С/П (3 сек) N_4 С/П (3 сек) N_5 С/П (получим индекс I).

При неизменных баллах для расчета нового индекса начинать с п. б., а при неизменном N - с пункта в.

Контрольный пример

β_i	*1*	*0.5*	*0*	*-0.5*	*-1*	
N_i	*20*	*20*	*30*	*10*	*10*	*N = 90*

Результат: I = 0,1666666

Ниже приводится программа расчета евклидовых расстояний между векторами, характеризующими социологические объекты (респондентов, групп респондентов). Такого рода расстояния используются для определения близости объектов в некотором пространстве признаков (типичный пример - разбиение массива опрошенных на некоторые типы в соответствии с содержательными гипотезами и проверка того, насколько удачно введены типы, путем расчета расстояний между ними). Эта программа представляет интерес тем, что результатом счета является не одно число, а матрица расстояний 4 × 4 (т.е. 6 чисел, так как матрица симметрична).

[245]

Программа расчета матрицы расстояний между 4-мя векторами

Представление данных

Вектора:

$X = \{x_1, x_2, ..., x_n\}$

$Y = \{y_1, y_2, ..., y_n\}$

$Z = \{z_1, z_2, ..., z_n\}$

$U = \{u_1, u_2, ..., u_n\}$где u - число признаков, x_i, y_i, z_i, u_i — значение i-го признака для векторов X, Y, Z, U соответственно.

Результаты

Матрица расстояний M между векторами X, Y, Z, U:

$$M = \begin{bmatrix} d_{XX} & d_{XY} & d_{XZ} & d_{XU} \\ d_{YX} & d_{YY} & d_{YZ} & d_{YU} \\ d_{ZX} & d_{ZY} & d_{ZZ} & d_{ZU} \\ d_{UX} & d_{UY} & d_{UZ} & d_{UU} \end{bmatrix}$$

Ввод программы

В/О	C_x	$P_,$	$P_,$	ПП	$F5$	$F4$	ПП	$P_,$	$F8$	$P_,$	P
P	$P6$	C_x	$C/П$	FC_x	$P8$	$P8$	FC_x	$P6$	—	$P6$	PP
$PП$	$P_,$	$P_,$	$F2$	$F4$	ПП	ПП	$F4$	БП	Fx^2	В/О	
	C_x	C_x	$P7$	$P8$	FC_x	FC_x	$P7$	$F2$	↑		
	$P_,$	$P_,$	$F3$	ПП	$F3$	$F5$	ПП	$F7$	$F6$		
	C_x	C_x	$P8$	FC_x	$P7$	$P8$	FC_x	↑	+		

Вычисление

1. В/О С/П (4 сек).

2. x_i Р2 y_i Р3 z_i Р4 u_i Р5 С/П (i = 1, 2, ..., n; 18 сек)

3. $F\sqrt{}$ (получим расстояние d_{YZ}), Р, $F\sqrt{}$ (расстояние d_{XZ}), Р, $F\sqrt{}$ (расстояние d_{XU}), Р, $F\sqrt{}$ (d_{YZ}), Р, $F\sqrt{}$ (d_{YU}) Р, $F\sqrt{}$ (d_{ZU})

Расстояния главной диагонали равны нулю, остальные расстояния получаем из соображений симметрии $(d_{ij} = d_{ji})$

Контрольный пример

Исходные данные:

$X = \{1.7.3\}$

$Y = \{2.4.4\}$

$Z = \{1.1.3\}$

$U = \{5.5.2\}$

[246]

Результаты после округления

M=

0	3.317	6.000	4.582
3.317	0	3.317	3.742
6.000	3.317	0	5.742
4.582	3.742	5.745	0

Формулы расчета

$$d_{XY} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Остальные расстояния вычисляются аналогично.

Приведенные программы дают достаточно полное представление о возможностях ПК ЭВМ. Мы полагаем, что использование ПК ЭВМ для вторичной обработки социологической информации, особенно для расчета уровня значимости различий, целесообразно при любых формах организации обработки, в том числе и при наличии собственного ВЦ с диалоговым режимом работы. Дальнейшее совершенствование ПК ЭВМ (увеличение объема памяти и быстродействия, и фиксирование программ на магнитных пластинках, избавляющие от необходимости вводить при включении программу заново) существенно расширит сферу применения микрокалькуляторов. Простота обращения, не предполагающая специальной подготовки, а также возможность использования их в качестве обычного микрокалькулятора, позволяет предположить, что программируемые клавишные ЭВМ станут для социологов таким же настольным средством анализа информации, как микрокалькуляторы.

[247]

Приложение 1.

О ВЕРОЯТНОСТИ

Создание индуктивной логики сопутствовало изучению простейших форм движения материи – макротел, земных и небесных. Как правило, предполагалось, что причинные зависимости здесь – однозначные, а случайность можно игнорировать. Это значит, что при некоторых условиях данное явление всегда происходит.

Однако при изучении микромира, а также социальных закономерностей исследователи столкнулись с явлениями, которые при заданных условиях могут происходить, а могут и не происходить. Здесь необходимость пробивается через случайность, которой в принципе пренебрегать нельзя, а изучаемые связи характеризуются множественностью причин и следствий. Для их описания возникла необходимость перехода к новым, статистическим методам,

Выводы статистики не обладают безусловностью, вскрываемые закономерности проявляются не в каждом отдельном случае, а лишь в массе однотипных явлений. Статистические закономерности – это закономерности массовых явлений. Выражаются они с помощью специальных категорий статистической науки - таких как средние, меры вариации, показатели тесноты связи и т.д. В массовых явлениях взаимосвязи проявляются в виде тенденций, в виде изменения средних величин при взаимном погашении случайностей.

Противопоставление статистических закономерностей классическим носит условный характер. Так, справедливо заметил Н. Винер, в работах основателя классической физики И. Ньютона «содержалась важная статистическая оговорка»[2], недооцененная современниками и последователями великого ученого. На это же обстоятельство указывал еще М. Дробиш, отмечавший статистический характер даже законов Кеплера: они «...определяют только *средние* пути движения планет, от которых последние постоянно уклоняются то в одну, то в другую сторону»[3].

Логической основой статистических методов исследования является вероятность. Чтобы в самых общих чертах ознакомиться с этим фундаментальным понятием современной науки, обратимся к рассмотрению событий. Предположим, что имеется некоторый, вполне определенный комплекс условий. В этих условиях осуществляется серия испытаний, результаты которых суть некоторые события.

[248]

[2] *Винер Н.* Кибернетика и общество. М., 1958, с. 24.
[3] *Дробиш М*. Нравственная статистика. СПб., 1867, с. 7.

Если событие неизбежно происходит при данных условиях, оно называется *достоверным.* Например, при нормальном давлении – 760 мм ртутного столба - и температуре 100°С химически чистая вода (этот перечень – комплекс условий) всегда закипает (достоверное событие).

Существуют события, которые при заданных условиях могут произойти, а могут и не произойти. Такие события мы будем называть *случайными.*

Классические примеры случайных событий - выпадение герба при бросании монеты, извлечение туза из колоды карт, выпадение шестерки при бросании кости и т.д. Другие примеры – поступление выпускника средней школы в вуз, удовлетворенность работника, выбираемого наугад из некоторого коллектива, своей специальностью и т.п.

Если при данном комплексе условий некоторое событие заведомо не может произойти, оно называется *невозможным.* Например, затвердевание воды (событие) при реализации выше упомянутых условий. Разумеется, при других условиях это же событие перестает быть невозможным. Таким образом, определение события соотносится с некоторым комплексом условий.

Если данный комплекс условий реализуется многократно, то становится возможным не только констатировать случайность события А, но и количественно оценить возможность его появления,

Длительные опытные наблюдения над появлением случайного события (при неизменном комплексе условий) показывают, что для довольно широкого круга явлений число появлений события подчиняется устойчивым закономерностям.

Пусть n - общее число испытаний, а m - число тех из них, которые завершились появлением случайного события А. Если проводить большое число независимых испытаний в неизменных условиях, то, как показывает опыт, частость $v = \frac{m}{n}$ незначительно отклоняется от некоторого числа p, которое, по определению, называется вероятностью А. Чем больше проведено испытаний, тем ближе частость к вероятности.

Вероятность, таким образом, определяется как средняя частость при большом числе испытаний или точнее: $p = \lim_{n\to\infty} \frac{m}{n}$ Например, при бросании монеты Бюффон для n = 4040 получил частость выпадения герба v = 0,5080, Пирсон для n = 12000 v = 0,5016, а для n = 24000 v = 0,5005. В качестве v данном случае принимается 0,5. Любопытно, что впервые устойчивость частости была обнаружена при изучении демографических явлений на большой статистике.

Впоследствии было также установлено, что распределение по росту, ширине грудной клетки, длине ступни людей определенного пола, возраста и национальности подчиняется устойчивой закономерности. Любопытный пример приводит в своей книге «Опыт философии теории вероятностей» Лаплас. Изучая закономерности рождения мальчиков и девочек, он обнаружил, что для статистических материалов Лондона, Берлина, Петербурга, Франции (в целом) относительные частоты рождения мальчиков в течение десятилетий колеблются около: $\frac{22}{43} \approx 0.512$ Для самого Парижа, однако, получалась несколько меньшая цифра: $\frac{25}{49} \approx 0.510$. Лаплас заинтересовался этим различием (в две тысячных!) и обнаружил, что в общее число рождений во французской столице включаются подкидыши; выяснилось также, что окрестное сельское население

[249]

преимущественно подкидывает девочек, что и исказило картину. Исключив подкидышей, Лаплас и для Парижа получил 22/43. Очевидно, это число и есть вероятность рождения мальчика.

Из определения вероятности события А следует, что $0 \le P(A) \le 1$, причем 0 соответствует невозможному событию, 1 – достоверному.

Итак, согласно нашей схеме, в одних и тех же условиях можно провести неограниченное число испытаний, в каждом из которых событие А может появиться, а может и не появиться; в результате большого числа испытаний устанавливается, что частота появлений события А для каждой большой группы испытаний мало отличается от некоторой постоянной величины, значение этой постоянной, по определению, называется *статистической вероятностью.*

При статистическом определении для нахождения вероятности необходимо проведение большого числа испытаний (Бюффон, Пирсон, Лаплас). В ряде случаев оказывается возможным дать априорное определение вероятности события, т.е. без фактических испытаний.

Для введения классического определения вероятности необходимо познакомиться с некоторыми понятиями. События бывают составными (например, выпадение при бросании кости не менее 5 очков) и элементарными (выпадение 5 очков). Элементарные события нельзя разложить на более простые, а составные можно разложить на элементарные. В ранее рассмотренном примере («не менее 5 очков») событие разложимо на две элементарных: выпадение 5 очков, выпадение 6 очков. Кстати, оба этих элементарных события благоприятствуют нашему событию. Обратим внимание и на то, что в случае бросания кости все шесть элементарных событий (выпадение каждой из шести граней) равновозможны, если кость правильная, и, естественно, несовместны, т.е. никакие два из них не могут появиться при одном испытании.

По определению, *классической вероятностью* события А называется отношение числа равновозможных элементарных событий, благоприятствующих событию А, к общему числу равновозможных элементарных событий:

В нашем примере: $P(A) = \frac{2}{6} = \frac{1}{3}$.

Пусть комплекс условий (испытание) состоит в подбрасывании двух идеальных костей. Какова вероятность того, что шестерка выпадает два раза (А)? Один раз (В)? Хотя бы один раз (С)?

Теперь имеется всего 36 элементарных равновозможных событий: каждая из граней одной кости выпадает с каждой гранью другой.

Событию А - «выпадает два раза» - благоприятствует только одно элементарное событие, следовательно, $P(A) = 36$. Событию В благоприятствуют такие элементарные: на первой кости шестерка, а на другой не шестерка (их 5), на первой не шестерка, на второй - шестерка (их тоже 5), т.е. $P(B) = \frac{10}{36} = \frac{5}{18}$. Событию С благоприятствуют все те, которые благоприятствуют В плюс еще одно: шестерка на первой и на второй кисти. Теперь $P(C) = \frac{10+1}{36} = \frac{11}{36}$.

Отметим, что если есть два независимых события А и В (т.е. два таких события, что осуществление одного из них никак не сказывается на осуществлении другого), то вероятность совместного осуществления А и В равна P (А) • P (В), Так, в рассмотренном примере выпадение шестерки на одной кости никак не влияет на выпадение той или иной

грани на другой кости, а поскольку вероятность выпадения шестерки на одной кости равна $\frac{1}{6}$ то выпадение шестерки на двух костях одновременно равно $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$.

Как видим, этот результат совпадает с полученным ранее из других соображений.

Упражнение 98.

1. Подбрасываются две монеты. Найти вероятность выпадения герба хотя бы на одной из них. Ответ: $\frac{3}{4}$.

2. То же в случае трех монет. Ответ: $\frac{7}{8}$.

3. Подбрасываются 3 монеты. Какова вероятность выпадения герба на всех монетах. Ответ: $\frac{1}{8}$.

4. В партии из 100 ламп 6 бракованных. Какова вероятность, что из двух ламп, взятых на испытание, обе окажутся неисправными?

Указание: найти вероятность, что 1-я лампа бракованная; после проверки 1-й лампы остается 99 ламп - найти вероятность того, что 2-я лампа бракованная, воспользоваться формулой перемножения вероятностей независимых событий. Ответ: $\frac{6\times5}{100\times99}\approx0.003$

Недостатком классического определения, очевидно, является дискретность пространства элементарных событий. Этот недостаток устраняется в так называемом геометрическом понятии вероятности. Пусть на плоскости имеется некоторая область *G*, внутри которой содержится область *g*. Комплекс условий заключается в бросании наудачу точки. Чему равна вероятность попадания точки в *g*?

Пространство элементарных событий теперь непрерывно, оно состоит из бесконечного числа точек. Совокупность точек области *g* образует множество элементарных событий, благоприятствующих обсуждаемому. Так как условие равновозможности сохраняется и здесь, то естественно определить вероятность события как отношение мер областей (в данном случае - площадей, но это может быть отношение длин отрезков, объемов и т.д.)

Пространство элементарных событий имеет нулевую вероятность (мера точки есть нуль). В то же время любое из элементарных событий является возможным. Таким образом, существуют события, осуществление которых на опыте возможно, хотя они имеют нулевую вероятность.

Как уже отмечалось, геометрическое определение, как и классическое, требует выполнения условия *равновозможности.* В этом его существенный недостаток. Этот недостаток преодолевается в так называемом аксиоматическом определении, изложение которого выходит из рамки данной книги[4].

Приложение 2.

СУММЫ И НЕКОТОРЫЕ ЗАДАЧИ НА СУММИРОВАНИЕ

О суммах. Для записи суммы ряда чисел используется греческая прописная буква $\sum$ (сигма), Например, $A_1+A_2+A_3+A_4$ записывается $\sum_{i=1}^{4}A_i$. Если у нас n слагаемых, то их записывают следующим

[251]

[4] Любознательного читателя мы отсылаем к книге: *Колмогоров А. Н.* Основные понятия теории вероятностей. М., 1974; См. также: *Пугачев* В. С. Теория вероятности и математическая статистика. М., 1979.

образом $A_1 + A_2 + ...A_n = \sum_{i=1}^{n} A_i$ и читают: «сумма A_i, где *i* изменяется дот 1 до *n*». Это обозначение существенно облегчает запись сумм.

Из определения $\sum$ следует, что

$$\sum_{i=1}^{n}(\alpha a_i + \beta b_i + \gamma) = \alpha\sum_{i=1}^{n} a_i + \beta\sum_{i=1}^{n} b_i + n\gamma$$

Действительно,

$$\sum_{i=1}^{n}(\alpha a_i + \beta b_i + \gamma) = (\alpha a_1 + \beta b_1 + \gamma) + (\alpha a_2 + \beta b_2 + \gamma) +$$

$$+ ... + (\alpha a_n + \beta b_n + \gamma) = \alpha(a_1 + a_2 + ... + a_n) +$$

$$+ \beta(b_1 + b_2 + ... + b_n) + n\gamma = \alpha\sum_{i=1}^{n} a_i + \beta\sum_{i=1}^{n} b_i + n\gamma$$

Отметим, что результат суммирования не зависит от того, как обозначен индекс суммирования:

$$\sum_{i=1}^{n} A_i = \sum_{i=1}^{n} A_S = A_1 + A_2 + ... + A_n$$

За знак суммы по i , например, можно выносить любое выражение, не содержащее i , даже если оно содержит другие индексы или суммы.

Например,

$$\sum_{i=1}^{n} A_i B_j = B_j \sum_{i=1}^{n} A_i$$

$$\sum_{i=1}^{n}\left(A_i \sum_{j=1}^{k} C_j B_j \right) = \left(\sum_{j=1}^{k} C_j B_j \right) \sum_{i=1}^{n} A_i$$

О суммировании степеней натуральных чисел.

Рассмотрим

$$S_n^{(k)} = 1^k + 2^k + ... + n^k = \sum_{i=1}^{n} i^k$$

При $k = 1$

$$S_n^{(1)} = 1 + 2 + ... + n = \frac{1+n}{2} n$$

— обычная сумма членов арифметической прогрессии.

Особый интерес представляет для нас случай *k*=2:

$$S_n^{(2)} = 1^2 + 2^2 + ... + n^2$$

Покажем как найти S_n^2

Рассмотрим

$$(n+1)^3 - n^3 = 3n^2 + 3n + 1$$

[252]

$$n^3-(n-1)^3=3(n-1)^2+3(n-1)+1$$

.....

$$3^3-2^3=3\cdot 2^2+3\cdot 2+1$$

$$2^3-1^3=3\cdot 1^2+3\cdot 1+1$$

Просуммируем почленно все равенства:

$$(n+1)^3-1=3\cdot S_n^{(2)}+3\cdot S_n^{(1)}+n$$

Так как $S_n^{(1)}$ нам известно, то последнее уравнение содержит лишь одну неизвестную величину $S_n^{(2)}$. Решая его, получим после простых преобразований:

$$S_n^{(2)}=\frac{n(n+1)(2n+1)}{6}.$$

Справедливость этого утверждения может быть показана и с помощью метода математической индукции.

Найдем $S_n^{(3)}$. Для этого можно использовать соотношение

$$(n+1)^4-1=4\cdot n^3+6\cdot n^2+4n+1,$$

в справедливости которого просто убедиться. Записывая его *n* раз, как это было выше сделано при нахождении $S_n^{(2)}$, получим после почленного суммирования

$$(n+1)^4-1=4\cdot S_n^{(3)}+6\cdot S_n^{(2)}+4S_n^{(1)}+n.$$

Отсюда

$$S_n^{(3)}=\left[\frac{n(n+1)}{2}\right]^2.$$

При рассмотрении коэффициента ранговой корреляции Спирмена необходимо знание суммы

$$\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2=\sum_{i=1}^{n}i^2-(n-1)\sum_{i=1}^{n}i+\frac{n(n-1)^2}{4}=$$

$$=\frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)^2}{2}+\frac{n(n+1)^2}{4}=\frac{n^3-n}{12}.$$

а также суммы квадратов *n* нечетных чисел:

$$\overline{S}_n^{(2)}=1^2+3^2+5^2+...+(2n-1)^2$$

Покажем, что $\overline{S}_1^2=\frac{n(4n^2-1}{3}$ с помощью метода математической индукции.

Для $n=1$ $\overline{S}_1^2=1$ — по определению и $\frac{1\cdot 3}{3}+1$ – по формуле.

(Для $n=2:\overline{S}_2^2=1+3=10$ - по определению и $\frac{2\cdot 15}{3}=10$ по формуле.

[253]

Пусть $\overline{S}_k^2 = \frac{k(4k^2-1)}{3}$, покажем, что при этом $\overline{S}_{k+1}^{(2)} = \frac{(k+1)(4k^2+8k+3)}{3}$

$$\overline{S}_{k+1}^{(2)} = \overline{S}_k^{(2)} + (2k+1)^2 = \frac{k(4k^2-1)}{3} +$$

$$+ (2k+1)^2 = \frac{(k+1)(4k^2+8k+3)}{3}.$$

Упражнение 99. Найти сумму квадратов *n* четных чисел

$$\overline{=}^{(2)}_{S_n} = 2^2 + 4^2 + 6^2 + \ldots + (2n)^2$$

Ответ: $\frac{2}{3}n(n+1)(2n+1)$

Упражнение 100. Показать, что $1^3 + 3^3 + \ldots + (2n+1)^3 = (n+1)^2(2n^2+4n+1)$

В некоторых случаях величины могут оказаться снабженными двумя индексами. Например, в корреляционной таблице 15 (§ 1 главы II) N_{ij} - число индивидов, у которых некоторый признак X принимает значение x_i а другой признак Y значение y_i. Очевидно, $\sum_{i=1}^{k} N_{ij}$, где k - число значений признака X, дает сумму элементов j-ого столбца таблицы, т.е. число индивидов, у которых $Y=y_i$ (независимо от того, каково значение признака X).

Аналогично $\sum_{j=1}^{l} N_{ij}$, где l - число значений признака Y, дает сумму элементов i-ой строки матрицы, т.е. число индивидов, у которых $X=x_i$ (независимо от того, каково значение признака Y).

Суммируя по обоим индексам, мы, очевидно, получим общее число наблюдений (количество индивидов обследуемой общности), т.е.

$$N = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}$$

Приложение 3.

СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Приведенные ниже таблицы частично заимствованы из других изданий, частично рассчитаны на «Электронике БЗ - 21» по составленным авторами программам. При подборе таблиц мы руководствовались, во-первых, соображениями удобства (приводимые в некоторых изданиях таблицы требуют иногда пересчета при пользовании[5] - здесь этот недостаток устранен); во-вторых, соображениями соответствия таблиц рас-

[254]

[5] См., например: Статистические методы анализа информации в социологических исследованиях, М., 1979, табл. А на с. 299.

пространенным в социологических исследованиях формам представления и количественным характеристикам информации (например, таблица χ^2 часто приводится в статистической литературе[6] лишь для числа степеней свободы, не превышающих 30, в то время как в социологических исследованиях нередки таблицы 7 × 10, 10 × 10 и т.д. с 50 - 80 и даже большим числом степеней свободы). Приведенные ниже таблицы составлены, как правило, для числа степеней свободы или объема выборки от 1 до 1000 и для уровней значимости 5%, 1% и 0,1%. Думается, что такая стандартизация облегчит пользование таблицами.

Исходя из высказанных в гл. V соображений о большей опасности ошибок I рода, чем II рода, *все таблицы приводятся для двустороннего критерия* (что уменьшает вероятность ошибок I рода и увеличивает вероятность ошибок II рода). Из этих же соображений рекомендуем читателю в случае, если в таблице не приведено критическое значение для полученной им статистической характеристики, брать с некоторым запасом ближайшее большее критическое значение (или же прибегнуть к интерполяции).

Отметим также, что при работе над таблицами были обнаружены опечатки в некоторых изданиях[7].

Таблица А

Нормальное распределение[8]

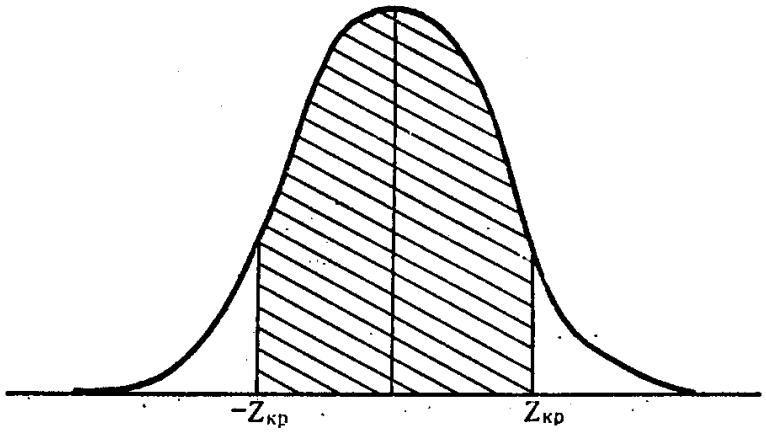


Рис. 31. Доли площади (Р) под нормальной кривой в пределах от $-z_{кр}$ до $+z_{кр}$

$z_{кр}$	P	$z_{кр}$	P	$z_{кр}$	P	$z_{кр}$	P
0,01	0,00000	02	01596	04	03191	06	04784
01	00798	03	02393	05	03988	07	05581

[255]

[6] Там же, табл. Б на с. 300, 301.

[7] Статистические методы..., с. 300, последняя строка: вместо 7,251 должно быть 6,25l, а вместо 6,815 должно быть 7,815; *Венецкий И. Г., Венецкая В. И.* Основные математико-статистические понятия и формулы в экономическом анализе. М., 1979, с. 413, последняя строка: вместо 1,77 должно быть 2,77.

[8] *Венецкий И. Г., Венецкая В. И.* Основные математико-статистические понятия..., с, 402 - 404.

Продолжение табл. А

$z_{кр}$	P	$z_{кр}$	P	$z_{кр}$	P	$z_{кр}$	P
08	06376	57	43132	06	71086	1,55	0,87886
09	07171	58	43809	07	71538	56	88124
0,10	0,07966	59	44481	08	71986	57	88358
11	08759	0,60	0,45149	09	72429	58	88589
12	09552	61	45814	3,10	0,72867	59	88817
13	10348	62	46474	11	73300	1,60	0,89040
14	11134	63	47131	12	73729	61	89260
15	11924	64	47783	13	74152	62	89477
16	12712	65	48431	14	74571	63	89690
17	13499	66	49075	15	74986	64	89899
18	14285	67	49714	16	75395	65	90106
19	15069	68	50350	17	75800	66	90309
0,20	0,15852	69	50981	18	76200	67	90508
21	16633	0,70	0,51607	19	76595	68	90704
22	17413	71	52230	3,20	0,76986	69	90897
23	18191	72	52848	21	77372	1,70	0,91087
24	18967	73	53461	22	77754	71	91273
25	19741	74	54070	23	78130	72	91457
26	20514	75	54675	24	78502	73	91637
27	21284	76	55275	25	78870	74	9144
28	22052	77	55870	26	79233	75	91588
29	22818	78	56461	27	79592	76	92159
0,30	0,23582	79	57047	28	79945	77	92327
31	24344	0,80	0,57629	29	80295	78	92492
32	25103	81	58206	3,30	0,80640	79	92655
33	25860	82	58778	31	80980	1,80	0,92814
34	26614	83	59346	32	81316	81	92970
35	27366	84	59909	33	81648	82	93124
36	28115	85	60468	34	81975	83	93275
37	28862	86	61021	35	82298	84	93423
38	29605	87	61570	36	82617	'85	93569
39	30346	88	62114	37	82931	86	93711
0,40	0,31084	89	62653	38	83241	87	93852
41	31819	0,90	0,63188	39	83547	88	93989
42	32552	91	63718	3,40	0,83849	89	44 124
43	33280	92	64243	41	84146	1,90	0,94257
44	34006	93	64763	42	84439	91	94387
45	34729	94	65278	43	84728	92	94514
46	35448	95	65789	44	85013	93	94639
47	36164	96	66294	45	85294	94	94762
48	36877	97	66795	46	85571	95	94882
49	37587	98	67291	47	85844	96	95000
0,50	0,38292	99	67783	48	86113	97	95116
51	38995	1,00	0,68269	49	86378	98	95230
52	39694	01	68750	3,505	0,86639	99	95341
53	40389	02	69227	51	86696	2,00	0,95450
54	41080	03	69699	52	87149	01	95557
55	41768	04	70166	53	87398	02	95662
56	42452	05	70628	54	87644	03	95764

Продолжение табл. А.

zкр	P	zкр	P	zкр	P	zкр	P
04	95865	53	98859	02	99747	51	99955
05	95964	54	98891	03	99755	52	99957
06	96060	55	98923	04	99763	53	99958
07	96155	56	98953	05	999771	54	99960
08	96247	57	98983	06	99779	55	99961
09	96338	58	99012	07	99786	56	99963
2,10	0,96427	59	99040	08	99793	57	99964
11	96514	2,60	0,99068	09	99800	58	99966
12	96599	61	99095	3,10	0,99806	59	99967
13	96683	62	99121	11	99813	3,60	0,99968
14	96765	63	99146	12	99819	61	99969
15	96844	64	99171	13	99825	62	99971
16	96923	65	99195	14	99831	63	99972
17	96999	66	99219	15	99837	64	99973
18	97074	67	99241	16	99842	3,65	0,99974
19	97148	68	99263	17	99848	66	99975
2,20	0,97219	69	99285	18	99853	67	99976
21	97289	2,70	0,99307	19	99858	68	99977
22	97358	71	99327	3,20	0,99863	69	99978
23	97425	72	99347	21	99867	3,70	0,99978
24	97491	73	99367	22	99872	71	99979
25	0,97555	74	99386	23	99876	72	99980
26	97618	75	99404	24	99880	73	99981
27	97679	76	99422	25	99855	74	99982
28	97739	77	99439	26	99889	75	99982
29	97789	78	99456	27	99892	76	99983
2,30	0,97855	79	99473	28	99896	77	99984
31	97911	2,80	0,99489	29	99900	78	99984
32	97966	81	99505	3,30	0,99903	79	99985
33	98019	82	99520	31	99907	3,80	0,99986
34	98072	83	99535	32	99910	81	99986
35	98123	84	99549	33	99913	82	99987
36	98172	85	99563	34	99916	83	9987
37	98221	86	99576	35	99919	84	99988
38	98269	87	99590	36	99922	85	99988
39	98315	88	99602	37	99925	86	99989
2,40	0,98360	89	99615	38	99928	87	99989
41	98405	2,90	0,99627	39	99930	88	99990
42	98448	91	99639	3,40	0,99933	89	99990
43	98490	92	99650	41	99935	3,90	0,99990
44	98531	93	99661	42	99937	91	99991
45	98571	94	99672	43	99940	92	99991
46	98611	2,95	0,99682	44	99942	93	99992
47	98649	96	99692	45	99944	94	99992
48	98686	97	99702	46	99946	95	99992
49	98723	98	99712	47	99948	96	99992
2,50	0,98758	99	99721	48	99950	97	99993
51	98793	3,00	0,99730	49	99952	98	99993
52	98826	01	99739	3,50	0,99953	99	99993

Таблица Б.

Значение χ^2_0 в зависимости от числа степеней свободы (f) и уровня значимости[9]

f	Уровень значимости, %			f	Уровень значимости, %			f	Уровень значимости, %		
	5	1	0,1		5	1	0,1		5	1	0,1
1	3,84	6,63	10,83	16	26,30	32,00	39,25	40	55,76	63,69	73,40
2	5,99	9,21	13,81	17	27,59	33,41	40,79	50	67,50	76,15	86,66
3	7,81	11,34	16,27	18	28,87	34,80	42,31	60	79,08	8838	99,61
4	9,49	13,28	18,46	19	30,14	36,19	43,82	70	90,53	100,42	112,33
5	11,07	15,09	20,52	20	31,41	37,57	45,31	80	101,88	112,33	124,84
6	12,59	16,81	22,46	21	32,67	38,93	46,80	90	113,15	124,12	137,21
7	14,07	18,47	24,32	22	33,92	40,29	48,27	100	124,34	135,81	149,45
8	15,51	20,09	26,12	23	35,17	41,63	49,73	200	233,99	249,44	267,50
9	16,92	21,67	27,88	24	36,41	42,98	51,18	300	341,40	359,91	—
10	18,31	23,21	29,59	25	37,65	44.31	52,62	400	447,63	468,72	—
11	19,67	24,72	31,26	26	38,88	45,64	54,05	500	553,13	576,49	—
12	21,03	26,22	32,91	27	40,11	46,96	55,48	600	658,09	683,52	—
13	22,37	27,69	34,53	28	41,34	48,28	56,89	700	762,66	789,97	—
14	23,68	29,14	36,12	29	42,56	49,59	58,30	800	866,91	895,98	—
15	25,00	30,58	37,70	30	43,77	50,89	59,70	1000	1074,68	1106,97	—

[258]

[9] Часть таблицы заимствована из кн.: *Закс Л.* Статистическое оценивание. М., 1976, с. 132—134, часть — из кн. *Оуэн Д. Б.* Сборник статистических таблиц. М., 1966, с. 49—55.

Таблица В.

Критические значения коэффициента ранговой корреляции ρ

Объем выборки	Уровень значимости, %			Объем выборки	Уровень значимости, %		
	5	1	0,1		5	1	0,1
6	0,829	1,000	_	25	0,398	0,510	0,618
7	0,745	0,893	1,000	30	0,362	0,466	0,570
8	0,691	0,857	0,952	35	0,333	0,429	0,534
9	0,683	0,817	0,917	40	0,311	0,402	0,501
10	0,636	0,782	0,891	45	0,294	0,380	0,475
11	0,618	0,754	0,867	50	0,279	0,361	0,450
12	0,580	0,727	0,823	60	0,254	0,330	0,415
13	0,555	0,698	0,801	70	0,235	0,306	0,385
14	0,534	0,675	0,793	80	0,220	0,286	0,361
15	0,518	0,654	0,760	90	0,207	0,270	0,341
16	0,500	0,632	0,741	100	0,196	0,257	0,324
17	0,485	0,615	0,734	150	0,160	0,209	0,265
18	0,472	0,598	0,709	200	0,139	0,182	0,231
19	0,458	0,583	0,694	500	0,087	0,115	0,148
20	0,445	0,568	0,679	1000	0,062	0,081	0,104

Критические значения при n < 12 рассчитаны нами по таблице точного распределения S *(Кендел М.* Ранговые корреляции. М., 1975, с. 188, 189). Значения при 12<*n*<30 при уровне значимости 5% и 1% заимствованы из книги: *Закс Л.* Статистическое оценивание. М., 1976, с. 369, остальные значения рассчитаны по формуле, полученной из (V,8,4).

Таблица Г.

Значимость τ при n ≤ 10

Вероятность того, что *S* для τ достигнет или превзойдет заданное значение (показаны только положительные величины; отрицательные определяются по симметрии)[10]

S	Значения *n*				S	Значения *n*		
	4	5	8	9		6	7	10
0	0,625	0,592	0,548	0,540	1	0,500	0,500	0,500
2	0,375	0,408	0,452	0,460	3	0,360	0,386	0,431
4	0,167	0,242	0,360	0,381	5	0,235	0,281	0,364
6	0,042	0,117	0,274	0,306	7	0,136	0,191	0,300
8		0,042	0,199	0,238	9	0,068	0,119	0,242
10		0,008	0,138	0,179	11	0,028	0,068	0,190
12			0,089	0,130	13	$0,0^{2}83$	0,035	0,146
14			0,054	0,090	15	$0,0^{2}14$	0,015	0,108
16			0,031	0,060	17		$0,0^{2}54$	0,078
18			0,016	0,038	19		$0,0^{2}14$	0,054

[259]

[10] *Кендэл М.* Ранговые корреляции. М., 1975.

Продолжение табл. Г.

S	Значения п				S	Значения п		
	4	5	8	9		6	7	10
20			0,027	0,022	21		0,0320	0,036
22			0,022	0,012	23			0,023
24			0,038	0,026	25			0,014
26			0,031	0,022	27			0,0283
28			0,042	0,021	29			0,0246
30				0,034	31			0,0223
32				0,031	33			0,0211
34				0,042	35			0,0347
36				0,052	37			0,0318
					39			0,0258
					41			0,0215
					43			0,0528
					45			0,0628

Примечание. Повторяющиеся нули заменены степенями: например, $0,0^{2}47$ означает 0,00047.

Таблица Д .

Критические значения коэффициента Кэндела τ при отсутствии объединенных рангов[11]

Объем выборки	Уровень значимости, %			Объем выборки	Уровень значимости, %		
	5	1	0,1		5	1	0,1
5	1,000	—	—	15	0,387	0,506	0,643
6	0,867	1,000		16	0,371	0,486	0,617
7	0,714	0,810	1,000	17	0,357	0,468	0,595
8	0,643	0,786	0,929	18	0,345	0,452	0,574
9	0,556	0,667	0,833	19	0,333	0,437	0,556
10	0,511	0,644	0,778	20	0,323	0,424	0,539
11	0,491	0,600	0,745	25	0,283	0,372	0,473
12	0,455	0,585	0,697	30	0,255	0,335	0,426
13	0,425	0,554	0,697	35	0,234	0,307	0,391
14	0,404	0,529	0,671	40	0,217	0,285	0,363

[260]

[11] Критические значения для $n \leq 12$ рассчитаны нами по таблице точного распределения τ *(Оуэн* Д. *Б.* Сборник статистических таблиц. М., 1966, с. 396—399). Остальные значения рассчитаны по формуле, полученной из (V,8,5).

Продолжение табл. Д

Объем выборки	Уровень значимости, %			Объем выборки	Уровень значимости, %		
	5	1	0,1		5	1	0,1
45	0,203	0,267	0,341	100	0,133	0,175	0,223
50	0,192	0,253	0,322	150	0,108	0,142	0,181
60	0,174	0,229	0,292	200	0,093	0,123	0,156
70	0,160	0,211	0,269	500	0,059	0,077	0,098
80	0,150	0,197	0,251	1000	0,041	0,054	0,056
90	0,141	0,185	0,236				

Поскольку при n>12 использовалось не точное распределение τ, а приближенное (см. глава IV), для n=13 мы получили значение 0,704, большее, чем для n=12. Поэтому в таблице проставлено то же значение, что и для n=12, т.е. 0,697.

Таблица Е.

Критические значения коэффициента корреляции r^{12}.

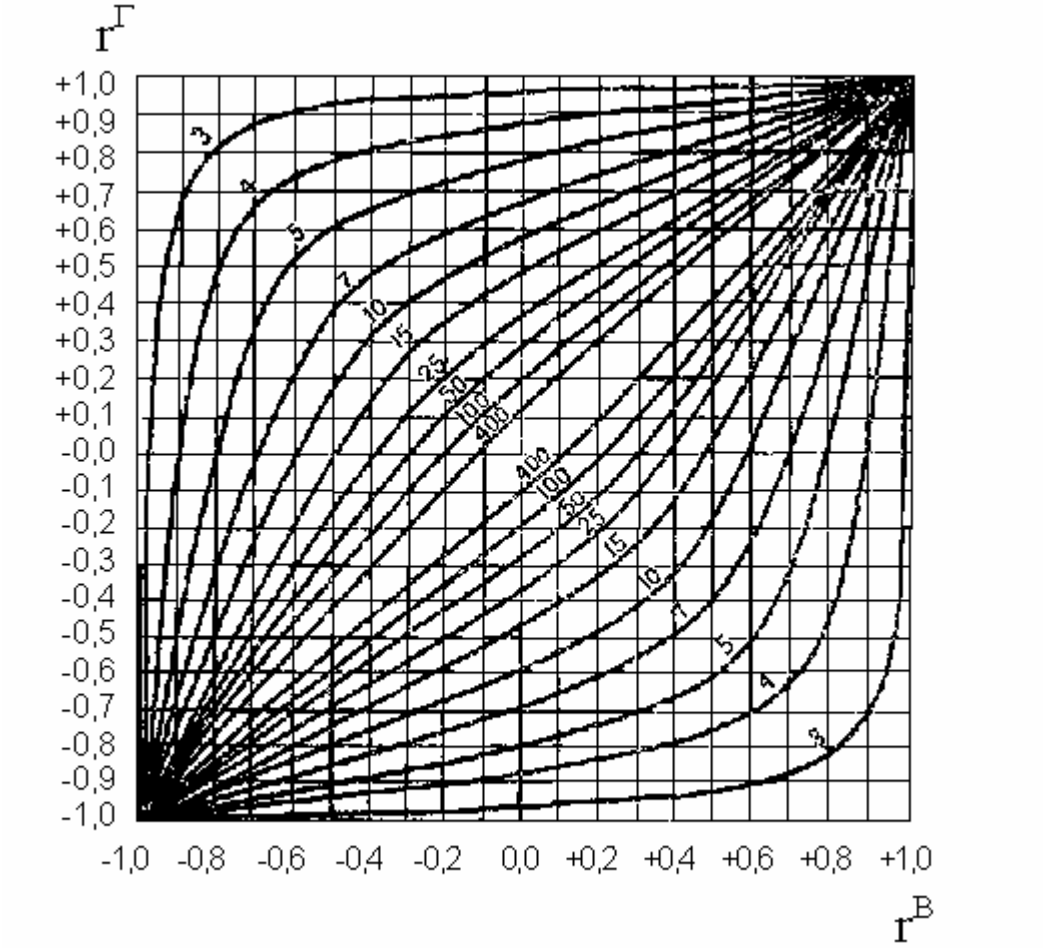
Объем выборки	Критические значения для 5%	1%	0.1%	Объем выборки	Критические значения для уровня 5%	1%	0,1%
3	0,99692	0,99988	_ _	20	0,4438	0,5614	0,6787
4	0,95000	0,99000	0,99900	21	0,4329	0,5487	0,6652
5	0,8783	0,95873	0,991 16	22	0,4227	0,5368	0,6524
6	0,8114	0,91720	0,97406	27	0,3809	0,4869	0,5974
7	0,7545	0,8745	0,95074	32	0,3494	0,4487	0,5541
8	0,7067	0,8343	0,92493	37	0,3246	0,4182	0,5189
9	0,6664	0,7977	0,8982	42	0,3044	0,3932	0,4896
10	0,6319	0,7646	0,8721	47	0,2875	0,3721	0,4648
11	0,6021	0,7348	0,8471	52	0,2732	0,3541	0,4433
12	0,5760	0,7079	0,8233	62	0,2500	0,3248	0,4078
13	0.5529	0,6835	0,8010	72	0,2319	0,3017	0,3799
14	0,5324	0,6614	0,7800	82	0,2172	0,2830	0,3568
15	0,5139	0,6411	0,7603	92	0,2050	0,2673	0,3375
16	0,4973	0,6226	0,7420	102	0,1946	0,2540	0,3211
17	0,4821	0,6055	0,7246	202	0,1381	0,1809	0,2299
18	0,4683	0,5897	0,7084	502	0,0875	0,1149	0,1464
19	0,4555	0,5751	0,6932	1002	0,0619	0,0813	0,1035

[261]

[12] *Джонсон Н., Лион Ф.* Статистика и планирование эксперимента в технике и науке. М., 1980, с. 560. Мы несколько дополнили заимствованную из этой книги таблицу.

Таблица Ж.

Номограмма[13] для определения доверительного интервала генерального коэффициента корреляции $r^{Г}$ по значению выборочного коэффициента корреляции $r^{В}$



Пояснение к номограмме. Отложив значение r^{B} на оси абсцисс, проводим перпендикуляр к ней до пересечения с двумя линиями, соответствующими объему выборки n; ордината первого пересечения даст нижнее значение, а ордината второго пересечения — верхнее значение доверительного интервала $r^{Г}$. Например, r^{B}=0,2 для n=50 даст примерно следующий доверительный интервал для $r^{Г}$: -0,10; 0,45.

[262]

[13] *Джонсон, Н., Лион Ф.* Статистика и планирование эксперимента..., с. 559.

Таблица З.

Номограммы[14] для определения доверительного интервала доли в генеральной совокупности ($\nu^{Г}$) по доле в выборке ($\nu^{В}$).

Пояснение к номограммам (см. с. 264)

Пользоваться номограммами аналогично предыдущему случаю.

Пример: пусть $\nu^{В} = 0{,}45$, тогда доверительный интервал для n = 250 при P = 0,95 равен приблизительно 0,38; 0,52, а при *P* = 0,99 0,36; 0,55.

Таблица И.

Критические значения критерия *t* Стьюдента в зависимости от числа степеней свободы (*f*) и уровня значимости[15]

df	Уровни значимости, %			df	Уровни значимости, %		
	5	1	0,1		5	1	0,1
1	12,71	63,60		21	2,08	2,83	3,82
2	4,30	9,93	31,60	22	2,07	2,82	3,79
3	3,18	5,84	12,94	23	2,07	2,81	3,77
4	2,78	4,60	8,61	24	2,06	2,80	3,75
5	2,57	4,03	6,86	25	2,06	2,79	3,73
6	2,45	3,71	5,96	26	2,06	2,78	3,71
7	2,37	3,50	5,41	27	2,05	2,77	3,69
8	2,31	3,36	5,04	28	2,05	2,76	3,67
9	2,26	3,25	4,78	29	2,04	2,76	3,66
10	2,23	3,17	4,59	30	2,04	2,75	3,65
11	2,20	3,11	4,44	40	2,02	2,70	3,55
12	2.18	3,06	4,32	50	2,01	2,68	3,50
13	2,16	3,01	4,22	60	2,00	2,66	3,46
14	2,15	2,98	4,14	80	1,99	2,64	3,42
15	2,13	2.95	4,07	100	1,98	2,63	3,39
16	2,12	2,92	4,02	120	1,98	2,62	3,37
17	2,11	2,90	3,97	200	1,97	2,60	3,34
18	2,10	2,88	3,92	500	1,96	2,59	3,31
19	2,09	2,86	3,88	∞	1,96	2,58	3.29
20	2,09	2,85	3,85				

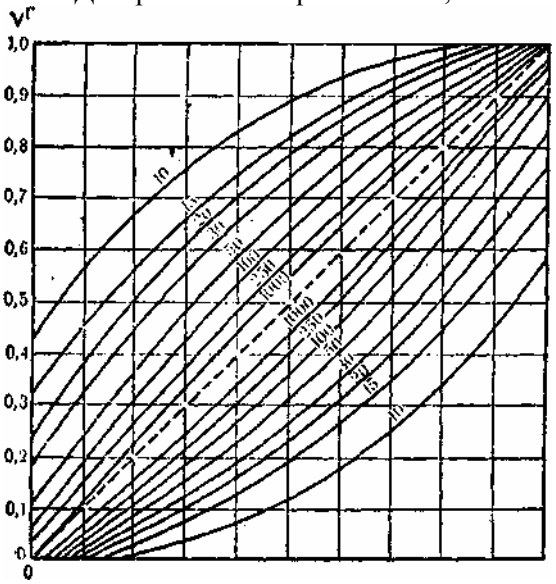
[263]

[14] Джонсон Н., Лион Ф. Статистика и планирование эксперимента..., с. 537, 538.

[15] Францевич Л. И. Обработка результатов биологических экспериментов на микро-ЭВМ «Электроника БЗ-21». Киев, 1979, с. 86.

Доверительная вероятность 0,95

Доверительная вероятность 0,99



[264]

Таблица К.

Преобразование Фишера $z = \frac{1}{2}\ln\frac{1+r}{1-r}$ [16]

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1104	0,1206	0,1307	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2985
0,3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5763	0,590!	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6466
0,99	2,6466	2,6996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

Слева в таблице размещены десятые, а сверху - сотые доли коэффициента корреляции. Например, для r = 0,42 ищем z на пересечении строки 0,4 и столбца 2. Получим 0,4477. Для обратного преобразования z в r находим внутри таблицы ближайшее к z число и по строке и столбцу определяем r. Например, для z = 0,76 ближайшее число 0,7582. Оно стоит в строке 0,6 и столбце 4, следовательно, r = 0,64.

[265]

[16] *Венецкий И. Г., Венецкая И. В.* Основные математико-статистические понятия и формулы в экономическом анализе, с. 423.

Таблица Л.

Значение F (критерий Фишера) при вероятностях: 0,95 (верхняя строка) и 0,99 (нижняя строка)[17]

f_2	Число степеней свободы для большей дисперсии f_1											
	1	3	5	7	9	11	14	20	30	60	120	∞
1	161	216	230	237	241	243	245	248	250	252	253	254
	4052	5403	5764	5923	6022	6032	6142	6208	6258	6313	6339	6366
3	10,13	9,28	9,01	8,88	8,81	8,76	8,71	8,66	8,62	8,57	8,55	8,53
	34,12	29,46	28,24	27,67	27,34	27,13	26,92	26,69	26,50	26,32	26,22	26,13
5	6,61	5,41	5,05	4,88	4,78	4,70	4,64	4,56	4,50	4,43	4,40	4,37
	16,26	12,06	10,97	10,45	10,15	9,96	9,77	9,55	9,38	9,20	9,11	9,02
7	5,59	4,35	3,97	3,79	3,63	3,60	3,52	3,44	3,38	3,30	3,27	3,23
	12,25	8,43	7,46	7,00	6,71	6,54	6,35	6,15	5,98	5,82	5,74	5,65
9	5,12	3,86	3,48	3,29	3,18	3,10	3,02	2,93	2,86	2,79	2,75	2,71
	10,56	6,99	6,06	5,62	5,35	5,13	5,00	4,80	4,64	4,48	4,40	4,31
11	4,84	3,59	3,20	3,01	2,90	2,82	2,74	2,65	2,57	2,49	2,45	2,40
	9,85	6,22	5,32	4,88	4,63	4,46	4,29	4,10	3,94	3,78	3,69	3,60
13	4,67	3,41	3,02	2,84	2,72	2,63	2,55	2,46	2,38	2,30	2,25	2,21
	9,07	5,74	4,86	4,44	4,19	4,02	3,85	3,67	3,51	3,34	3,25	3,17
15	4,54	3,29	2,90	2,70	2,59	2,51	2,43	2,33	2,25	2,16	2,11	2,10
	8,68	5,42	4,56	4,14	3,89	3,73	3,56	3,36	3.20	3,05	2,96	2,87
17	4,45	3,20	2,81	2,62	2,50	2,41	233	2,23	2,15	2,06	2,01	1,96
	8,40	5,18	4,34	3,93	3,68	3,52	3,35	3,16	3,00	2,83	2,75	2,65
19	4,38	3,13	2,74	2,55	2,43	2,34	2,26	2,15	2,07	1,98	1,93	1,88
	8,18	5,01	4,17	3,77	3,52	3,36	3,19	3,00	2,84	2,67	2,58	2,49
21	4,32	3,07	2,63	2,49	2,37	2 28	2,20	2,09	2,00	1,92	1,87	1,81
	8,02	4,87	4,04	3,65	3,40	3,24	3,07	2,88	2,72	2,55	2,46	2,36
23	4,28	3,03	2,64	2,45	2,32	2,24	2,14	2,04	1,96	1,86	1,81	1,76
	7,88	4,76	3,94	3,54	3,30	3,14	2,97	2,78	2,62	2,45	2,35	2,26
25	4,24	2,99	2,60	2,41	2,28	2,20	2,11	2,00	1,92	1,82	1,77	1,71
	7,77	4,68	3,86	3,46	3,21	3,03	2,89	2,70	2,54	2,36	2,27	2,17
27	4,21	2,06	2,57	2,37	2,25	2,16	2,08	1,97	1,88	1,78	1,73	1,67
	7,68	4,60	3,79	3,39	3,14	2,93	2,83	2,63	2,45	2,29	2,20	2,10
29	4,18	2,93	2,54	2,35	2,22	2,14	2.05	1,94	1,85	1,75	1,70	1,64
	7,60	4,54	3,73	3,33	3,08	2,92	2,77	2,57	2,41	2,23	2,14	2,03
30	4,17	2,92	2,53	2,33	2,21	2,13	2,04	1,93	1,84	1,74	1,68	1,62
	7,56	4,51	3,70	3,30	3,07	290	2,74	2,55	2,34	2,21	211	2,01
40	4,08	2,84	2,45	2,25	2,12	2,04	1,95	1,84	1,74	1,64	1,58	1,51
	7,31	4,31	3,51	3,12	2,89	2,73	2,56	2,37	2,20	2,02	1,98	1,80

[17] Фрагмент таблицы из книги: *Оуэн Д.Б.* Сборник статистических таблиц. М., 1966

60	4,00	2,76	2,37	2,17	2,04	1,95	1,86	1,75	1,65	1,53	1,48	1,39
	7,03	4,13	3,34	2,95	2,72	2,56	2,39	2,20	2,03	1,84	1,73	1,60
80	3,96	2,72	2,33	2,13	2,00	1,91	1,82	1,70	1,60	1,48	1,41	1,32
	6,96	4,04	3,26	2,87	2,64	2,48	2,31	2,12	1,94	1,75	1,63	1,49
120	3,92	2,68	2,29	2,09	1,96	1,87	1,77	1,66	1,55	1,43	1,35	1,25
	6,85	3,95	3,17	2,79	2,56	2,40	2,23	2,03	1,86	1,66	1,53	1,38
∞	3,84	2,60	2,21	2,01	1,88	1,79	1,69	1,57	1,46	1,32	1,22	1,00
	6,63	3,78	3,04	2,64	2,41	2,25	2,08	1,88	1,67	1,47	1,32	1,00

Таблица М.

Случайные числа[18]

63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
15474	45266	95270	79953	59367	83848	82396	10118	33211	59466
94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
42481	16213	97344	03721	16863	48767	03071	12059	25701	46670
23523	78317	73208	89837	63935	91416	26252	29663	05522	82562
04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
35963	15307	26398	09354	33351	35462	77974	50024	90103	39333
59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
46058	85236	01390	92286	77281	44077	93910	83647	70617	42941
32179	00597	87379	25241	05567	07007	86743	17157	85394	11833
69234	61406	20117	45204	15956	60000	18743	92423	97118	96333
19565	41430	01758	75379	40419	21585	66674	36806	84962	85207
45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
33185	16232	41941	50949	89435	43531	88695	41994	37548	73043
80951	00406	96382	70774	20151	23337	25016	25298	94624	61171
79752	49140	71961	28296	69861	02591	74852	20539	00387	59579
18633	32537	98145	06571	31010	24674	05455	61427	77938	91936
74029	43902	77557	32270	9779O	17119	52527	58021	80814	51748
54178	45611	80993	37143	05335	12969	56127	19255	36040	90324
11664	49883	52079	84827	59381	71539	09973	33440	88461	23356
48324	77928	31249	64710	02295	36870	32307	57546	15020	09994
69074	94138	87637	91976	35584	04401	10518	21615	01848	76938
09188	20097	32825	39527	04220	86304	83389	87374	64278	58CM4

[18] Фрагмент таблицы из кн.: Статистические методы анализа информации в социологических исследованиях, с. 305—308.

90045	85497	51981	50654	94938	81997	91870	76150	68476	64659
73189	50207	47677	26269	62290	64464	27124	67018	41361	82760
75768	76490	20971	87749	90429	12272	95375	05871	93823	43178
54016	44056	66281	31003	00632	27398	20714	53295	07706	17813
08353	69910	78542	42785	13661	53873	04618	97553	31223	08420
28306	03264	81333	10591	40510	07893	32604	60475	94119	01»40
53840	86233	81594	13628	51215	90290	28466	68795	77762	20791
91757	53741	61613	62269	50263	90212	55781	76514	83483	47055
89415	92694	00397	58391	12607	17646	48949	72306	94541	37408
77513	03820	86864	29901	68414	82774	51908	13980	72893	55507
19502	37174	69979	20288	55210	29773	74287	75251	65344	67215
21818	59313	93278	81757	05686	73156	07082	85046	31853	38452
51474	66499	68107	23621	94049	91345	42836	09191	08007	45449
99559	68331	62535	24170	69777	12830	74819	78142	43860	72834
33713	48007	93584	72869	51926	64721	58303	29822	93174	93942
85274	86893	11303	22970	28834	34137	73515	90400	71148	43643
84133	89640	44035	52165	73852	70091	61222	60561	62327	18423
56732	16234	17395	96131	10123	91622	85496	57560	81604	18880
65138	56806	87648	85261	34313	65361	45375	21069	85644	47277
38001	02176	81719	11711	71602	92937	74219	64049	65584	49698
37402	96397	01304	77586	56271	10086	47324	62605	40030	37438
97125	40348	87083	31417	21815	39250	75237	62047	15501	29578

[268]

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

N - объем генеральной совокупности
n - объем выборочной совокупности
X, Y .. - признаки (переменные)
x_i - i-е значение признака
N или N (x_i) - число индивидов, у которых признак X принимает значение x_i
$N(y_j)$ - число индивидов, у которых признак Y принимает
значение y_i
N_{ij} - число индивидов, у которых признак X принимает значение x_i, а признак Y значение y_i (эмпирическая частота)
N_{ij}^{0} - теоретическая частота
ν_i - частость (доля)
x_i' - левая граница 1-го интервала
x_i'' - правая граница ('-го интервала
I_i - ширина 1-го интервала
F_i - кумулятивная частота
f_i - кумулятивная частость
ρ_i -плотность
M или $\bar{x}$ - среднее арифметическое
M^{Γ} - генеральное среднее
M^{B} - выборочное среднее
Me - медиана
Mo - мода
G_N - среднее геометрическое
S_N - среднее квадратическое
H_N - среднее гармоническое
R - вариационный размах
D или σ^2 — дисперсия
σ - среднее квадратическое отклонение
C_υ - коэффициент вариации
s^2 - выборочная дисперсия (оценка σ^2)
α_k - нормированная мера вариации качественных признаков
Q_i – i-й квартиль $(\ i = \overline{1,3}\ )$

[269]

ΔQ — квартильное отклонение

D_i — i-й дециль $(\ i\ =\ \overline{1,9}\ )$

ΔD — децильное отклонение

χ^2 — критерий хи-квадрат Пирсона

f — число степеней свободы

φ^2 — средняя квадратичная сопряженность

C — коэффициент средней квадратической сопряженности

T — коэффициент Чупрова

T_c — коэффициент Крамера

Q - коэффициент ассоциации Юла для таблиц 2×2

Φ - коэффициент контингенции для таблиц 2×2

ρ- коэффициент ранговой корреляции Спирмена

τ - коэффициент ранговой корреляции Кендэла

r — коэффициент парной корреляции Пирсона — Браве

E — энтропия

ε — энтропийная мера дисперсии

$E_{y/x}$ — полная условная неопределенность Y — распределения

λ— энтропийная мера связи

$g,\ \gamma$ — коэффициенты Гудмана

δ — коэффициент близости разбиений

Δ — модульный коэффициент связи

d — коэффициент Сомерса

y_i — условное среднее

η — корреляционное отношение

$r_{\rho b}$ — ранговый бисериальный коэффициент корреляции

r_{rb} — точечный бисериальный коэффициент корреляции

z — критические значения нормального распределения

t— критические значения распределения Стьюдента

F — критические значения распределения Фишера

z — значения функции преобразования Фишера

$r_{01.2}$ — коэффициент частной корреляции

$R_{0.123}$ — коэффициент множественной корреляции

H_0 — нулевая гипотеза

H_1 — альтернативная гипотеза

q — уровень значимости

p — доверительная вероятность

[270]