РАЗДЕЛ IV. ИЗМЕРЕНИЕ СВЯЗИ МЕЖДУ ПРИЗНАКАМИ С ПОМОЩЬЮ МАТЕМАТИЧЕСКИХ МЕТОДОВ

4.1. Понятие статистической связи

Принцип взаимной сопряженности. Аксиомой науки, а также здравого смысла является убеждение в том, что ни одно событие в природе не возникает «вдруг», но всегда при вполне определенных, известных или неизвестных обстоятельствах. Событие, таким образом, никогда не следует рассматривать изолированно. Оно должно рассматриваться как результат совместных воздействий многих сил, каждая из которых влияет на наблюдаемый результат. Размер семьи, например, может зависеть от таких факторов, как продолжительность супружеской жизни, уровень дохода, степень занятости матери на работе вне дома, ее религиозные взгляды и т. д. Одни из этих факторов имеют тенденцию увеличивать, другие − уменьшать интенсивность изменения и даже препятствовать возникновению события. Так, религиозный фактор может благоприятствовать росту семьи, тогда как экономический фактор может притормозить эту тенденцию. Во всяком случае, нельзя предсказать размер семьи абсолютно точно, можно лишь связывать его с некоторыми факторами или переменными, с которыми сопряжен результат. Это знание основано, таким образом, на принципе взаимной сопряженности. Именно в соответствии с этим принципом и естественные, и социальные, и любые другие науки устанавливают свои методы и цели: 1) идентифицировать переменные (факторы), связанные с событием; 2) раскрыть способы взаимосвязи между факторами и событием; 3) измерить силу этой взаимосвязи.

Первая проблема заключается в выделении сопряженных факторов. Существует бесчисленное количество факторов, отношения между которыми столь запутаны, что «конечный» фактор при появлении события никогда не может быть получен и тем более измерен. Однако в бесконечных поисках определенности здравый смысл начинает конструировать образцы зависимостей, на базе которых он стремится охватить прошлое, понять настоящее и тем самым предвидеть будущее.

Виды зависимости. *Корреляция (корреляционная зависимость)* – статистическая взаимосвязь двух или нескольких случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Виды подобной взаимосвязи, конечно, пересматриваются в ходе каждодневного опыта, накапливаемого методом проб и ошибок. В этом процессе наблюдатель мысленно квантифицирует и суммирует свои наблюдения, во-первых, отмечая факторы, которые, как ему кажется, «производят» событие или просто связаны с ним, и, во-вторых, отмечая частоту, с которой он успешно предвидит или предсказывает. Он интуитивно

использует принцип корреляции.

Корреляционные связи различаются по форме, направлению и степени (силе) [6; 9; 14].

По форме корреляционная связь может быть прямолинейной или криволинейной. *Прямолинейной* может быть, например, связь между количеством посещенных занятий по курсу «Математические методы в социологии» во время семестра и количеством правильно решенных задач на экзамене по этому курсу. *Криволинейной* может быть, например, связь между уровнем мотивации и правильностью решения задач: при повышении мотивации (гарантированный, на определенных условиях «автомат» по экзамену) эффективность (правильность) выполнения задач сначала возрастает. Оптимальному уровню мотивации соответствует максимальная эффективность выполнения задач. Вполне возможно, что дальнейшему повышению мотивации может сопутствовать снижение эффективности.

По направлению корреляционная связь может быть положительной («прямой») и отрицательной («обратной»). При *положительной* прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого (см. *Рис. 4.1*). При *отрицательной* корреляции соотношения обратные. Если воспользоваться одним из уже рассматриваемых нами примеров, то в случае прямой связи мы имеем: «с возрастанием количества посещений занятий по курсу «Математические методы в социологии» эффективность решения задач возрастает. В случае же обратной связи: «с возрастанием количества посещений занятий по курсу «Математические методы в социологии» эффективность решения задач уменьшается» (что, конечно же, невероятно, и может рассматриваться только для иллюстрации направления связи!).

При положительной корреляции коэффициент корреляции имеет положительный знак, например r = +0,207, при отрицательной корреляции – отрицательный знак, например r = −0,207 [14, с. 44].

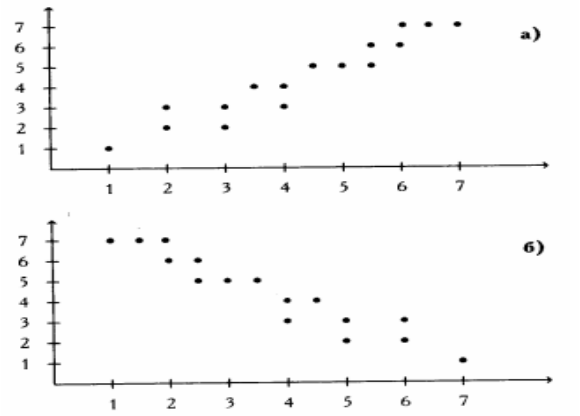

Рис. 4.1. Схема прямолинейных корреляционных связей: а) положительная (прямая) связь; б) отрицательная (обратная) связь

Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции.

Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции. Коэффициент корреляции – это величина, которая может варьировать в пределах от –1 до +1. В случае полной положительной корреляции этот коэффициент равен «+1», а при полной отрицательной – «–1».

На рисунке 4.2 представлено множество распределений двух признаков (X и Y) с соответствующими коэффициентами корреляций между этими признаками для каждого из распределений. Коэффициент корреляции отражает «зашумлённость» линейной зависимости (верхняя строка), но не описывает наклон линейной зависимости (средняя строка), и совсем не подходит для описания сложных, нелинейных зависимостей (нижняя строка). Для распределения, показанного в центре рисунка, коэффициент корреляции не определен, так как его дисперсия равна нулю.

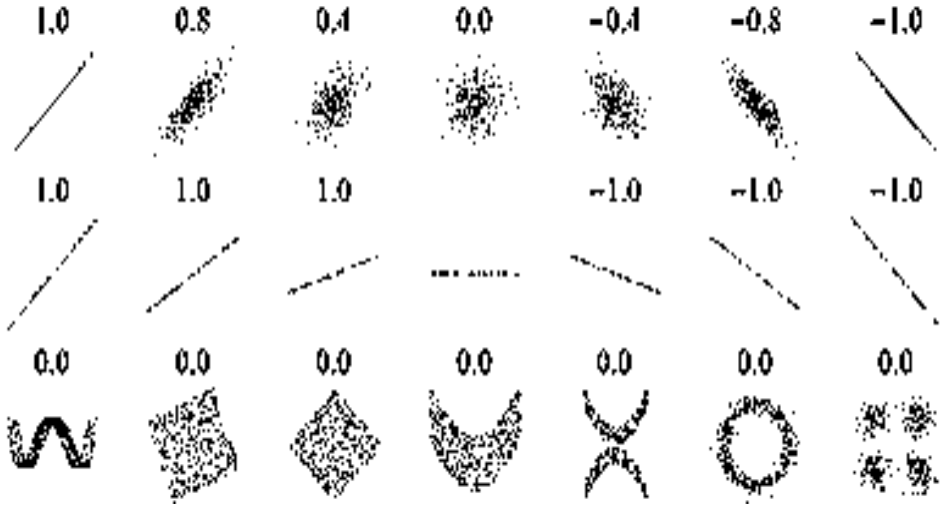

Рис. 4.2. Визуализация двумерных распределений по признакам X и Y, иллюстрирующая силу и направление связи между этими признаками

Метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными, носит название «*корреляционный анализ*». Корреляционный анализ тесно связан с *регрессионным анализом* (также часто можно встретить термин «*корреляционно-регрессионный анализ*»), который является более общим статистическим понятием [10]. С помощью регрессионного анализа определяют необходимость включения тех или иных факторов в уравнение множественной регрессии, а также оценивают полученное уравнение регрессии на соответствие выявленным связям, используя *коэффициент детерминации*. В целом, задачи регрессионного анализа лежат в сфере установления формы зависимости,

определения функции регрессии. А задачи корреляционного анализа сводятся к измерению тесноты связи между признаками. Величину, силу и направление связи между признаками показывают коэффициенты корреляции. Коэффициент корреляции определяется с помощь нахождения ковариации. *Ковариация* (или *корреляционный момент*) − числовая характеристика совместного распределения двух случайных величин равная математическому ожиданию произведения отклонений случайных величин от их математических ожиданий.

Для некоторых целей грубая субъективная оценка корреляции является вполне достаточной, но для научных целей желательны более точные измерения. Трудность заключается в сложности и разнообразии связей между социальными явлениями. Некоторые черты этой сложности могут быть сформулированы следующим образом: 1) каждое событие является результатом действия многочисленных факторов; 2) сила воздействия факторов изменяется по интенсивности; 3) она может быть направлена в одном или нескольких направлениях; 4) факторы находятся в постоянном взаимодействии; 5) они могут усиливать, противодействовать и уничтожать влияние друг друга.

Эта проблема, вероятно, менее остра в физических науках, чем в науках социальных. Физик посредством лабораторного контроля до определенной степени в состоянии выделять свой объект и манипулировать им. Он может тщательно повторить свои наблюдения в прежних условиях, в то время как социолог часто вынужден пользоваться данными, подобными необработанной руде, и собирать материал из разбросанных источников. Социолог, таким образом, вынужден использовать статистический контроль, так как лабораторный контроль ему недоступен.

Что является доказательством наличия связи? Как можно определить, что факторы находятся во взаимной связи? А, обнаружив связь, как можно определить степень или интенсивность этой связи? Вообще говоря, существуют два отличительных признака такого рода связи:

1) совместное появление качественных признаков;

2) параллельное изменение в двух или более рядах количественных переменных.

Во-первых, относительная частота, с которой появляются вместе определенные качественные признаки, может служить наиболее простым основанием для заключения об ассоциации (связи). В этом и заключается принцип совместного появления признаков. Статистические переменные, подобно людям, оцениваются обычно «в зависимости от общества, в котором они находятся». Например, если преступность более часто обнаруживается среди юношей, чем среди девушек, то заключают, что преступность ассоциируется с признаком «быть юношей». Сила этой связи будет изменяться в зависимости от других факторов, таких, как возраст юношей, характер преступления и многих других признаков. И все они будут затруднять статистическое применение этого простого, на первый взгляд, принципа. Следовательно, едва ли нужно еще раз повторять, что некоторая система

группирования и классификации необходима не только как средство установления наличия связи, но и как средство определения силы этой связи.

Во-вторых, если для двух рядов количественных данных изменение в одном из них соответствует с некоторой степенью устойчивости сравнимому изменению в другом ряду, то заключают, что они как-то связаны между собой и что существует связь между двумя рядами данных. Например, по мере того как падает доход, намечается тенденция к увеличению размера семьи; и если продолжить наблюдения на протяжении достаточно обширного ряда, то есть рассмотреть целый ряд семей различных размеров с различным доходом, то очевидность связи усиливается. Этот подход к оценке связи был назван принципом ковариации.

Способы и принципы измерения связи. Интерпретация коэффициентов корреляции. Техника измерения связи должна соответствовать: 1) природе данных; 2) числу взаимодействующих переменных; 3) видам зависимости между ними. Поэтому способы измерения связи будут различаться в зависимости от того, будут ли данные представлены в форме качественных признаков, которые просто перечислены, или в виде количественных измерений, и еще будет ли вид зависимости между переменными простым или сложным.

Используется две системы классификации корреляционных связей по их силе: общая и частная [14, с. 44−45].

Общая классификация **корреляционных связей***:*

1) сильная, или тесная при коэффициенте корреляции $К_{оэф} > 0{,}70$;

2) средняя при $0{,}50 < К_{оэф} < 0{,}69$;

3) умеренная при $0{,}30 < К_{оэф} < 0{,}49$;

4) слабая при $0{,}20 < К_{оэф} < 0{,}29$;

5) очень слабая при $К_{оэф} < 0{,}19$.

Для качественной оценки тесноты связи часто используют так называемую шкалу Чеддока, представленную в таблице 4.1.

Таблица 4.1

Оценка тесноты связи по шкале Чеддока

Количественная мера тесноты связи	Качественная характеристика силы связи
(0,1–0,3)	Слабая
[0,3–0,5)	Умеренная
[0,5–0,7)	Заметная
[0,7–0,9)	Высокая
(0,9–0,99)	Весьма высокая

Частная классификация *корреляционных связей:*

1) высокая значимая корреляция при r, соответствующем уровню статистической значимости $\rho \leq 0{,}01$;

2) значимая корреляция при r, соответствующем уровню статистической значимости $\rho \leq 0{,}05$;

3) тенденция достоверной связи при r, соответствующем уровню статистической значимости $\rho \leq 0{,}10$;

4) незначимая корреляция при r, не достигающем уровня статистической значимости.

Две эти классификации, − общая и частная, не совпадают. Первая ориентирована только на величину коэффициента корреляции, а вторая определяет, какого уровня значимости достигает данная величина коэффициента корреляции при данном объеме выборки. Чем больше объем выборки, тем меньшей величины коэффициента корреляции оказывается достаточно, чтобы корреляция была признана достоверной. В результате при малом объеме выборки может оказаться так, что сильная корреляция окажется недостоверной. В то же время при больших объемах выборки даже слабая корреляция может оказаться достоверной [14, с. 45].

Обычно принято ориентироваться на вторую классификацию, поскольку она учитывает объем выборки. Вместе с тем необходимо помнить, что сильная, или высокая, корреляция – это корреляция с коэффициентом $r > 0{,}70$, а не просто корреляция высокого уровня значимости.

Из многочисленных способов измерения связи рассмотрим только те из них, которые довольно просты в вычислении и наиболее часто используемы в социологической практике.

Формулы измерения связи могут быть удобно сгруппированы на основе двух рассмотренных выше принципов связи:

1) *совместного появления* (коэффициент взаимной сопряженности Пирсона (С), стандартные коэффициенты связи Чупрова и Крамера (Т и T_C);

2) *ковариации* (метод корреляции ранговых различий Спирмена (r_s) и Кенделла (τ), коэффициент корреляции Пирсона (r), корреляционное отношение (η), коэффициенты множественной ($R^2_{y(1\ldots k)}$) и частной корреляции ($r_{y1.2}$).

Все формулы и алгоритмы вычисления этих коэффициентов будут в подробностях и с применением реальных примеров рассмотрены нами, в подразделах 4.4−4.7 данного издания.

Коэффициенты рассчитываются с использованием стандартных программ и показывают меру взаимообусловленности в распределении частот появления соответствующих признаков. Один из признаков условно считается зависимым, другой – детерминирующим, однако заключение о наличии связи может дать только качественный анализ всей совокупности связей. Анализ коэффициентов связи позволяет:

- выделить факторы, статистический уровень влияния которых позволяет исключить их из дальнейшего анализа (гипотеза о наличии связи отрицается);

- проранжировать оставшиеся связи по уровню взаимной сопряженности с изучаемым процессом, при этом следует иметь в виду, что уровень взаимной сопряженности может определяться как влиянием данного фактора на процесс, так и взаимным изменением данного фактора и процесса под влиянием третьего фактора.

В заключение данного параграфа, приступая к более подробному и конкретному рассмотрению процедуры измерения взаимосвязей между признаками, хотелось бы обратить внимание на два следующих момента:

!!! во-первых, исходя из логических рассуждений, говоря о корреляции, не стоит отождествлять понятия «связь» и «зависимость». Если между признаками есть корреляционная связь – это еще не значит, что они взаимозависимы. Наличие связи между признаками может свидетельствовать не о зависимости этих признаков между собой, а о зависимости обоих этих признаков от какого-то третьего признака или сочетания признаков, не рассматриваемых в исследовании. Зависимость подразумевает влияние – любые согласованные изменения, которые могут объясняться сотнями причин [14, с. 40−41];

!!! во-вторых, как удачно отмечает автор одного из российских учебников по применению математических методов в психологии, Л. С. Титкова, корреляционные связи не могут рассматриваться как свидетельство причинно-следственной связи. Они свидетельствуют лишь о том, что изменениям одного признака, как правило, сопутствуют определенные изменения другого, но находится ли причина изменений в одном из признаков или она оказывается за пределами исследуемой пары признаков, нам неизвестно [14, с. 40−41].

4.2. Корреляционное поле

Корреляционная зависимость − связь между признаками, состоящая в том, что в зависимости от применения одного признака меняется величина другого. Как правило, при изучении взаимозависимости двух признаков различают: независимые признаки (факторные), которые чаще всего обозначаются − X; зависимый признак (результирующий) − Y.

В ходе корреляционного анализа необходимо узнать, как под влиянием факторных признаков изменяется результирующий, если он изменяется вообще и по какому-либо закону. Вспомогательным средством анализа выборочных данных является корреляционное поле. Если даны значения двух признаков X и Y, эти значения можно сопоставить путем отражения их в системе координат и нанесении на плоскость точек, соответствующих этим значениям и, при необходимости, соединения этих точек непрерывной линией. Расположение точек позволяет сделать предварительное заключение о характере и форме зависимости. Корреляционное поле относится к двумерной совокупности

данных точно так же, как гистограмма – к одномерной совокупности. Оно наглядно изображает распределение наблюдений в целом, тем самым позволяя получить грубую, но полезную оценку степени корреляции, прежде чем вычислить последнюю. Этот необходимый прием уже был использован для того, чтобы изобразить тенденцию временных рядов, однако детали построения поля еще не были описаны.

Чтобы проиллюстрировать более полно принципы построения корреляционного поля и его использования, возьмем в качестве примера таблицу 4.2, где представлена динамика развития сети высших учебных заведений Украины III–IV уровня аккредитации и численности студентов в них за 1992/2000 годы.

Таблица 4.2

Динамика развития сети высших учебных заведений Украины III–IV уровня аккредитации и численности студентов в них

Годы	*Количество высших учебных заведений Украины III–IV уровня аккредитации*	*Численность студентов в них (тыс.)*
[1992 – 1993)	158	718,8
[1993 – 1994)	159	680,7
[1994 – 1995)	232	645
[1995 – 1996)	255	617,7
[1996 – 1997)	274	595
[1997 – 1998)	280	526,4
[1998 – 1999)	298	503,7
[1999 – 2000)	313	503,7

Для того чтобы свести данные этой таблицы к корреляционному полю, сначала чертят горизонтальную и вертикальную оси, как и при построении гистограммы. Оси задаются как приблизительно равные по длине, если только нет достаточного основания для отступления от этого правила. Точно так же по обыкновению независимая переменная X располагается вдоль горизонтальной оси, а зависимая переменная Y – вдоль вертикальной.

Затем на осях устанавливаются шкалы таким образом, чтобы согласовать наблюдаемые области значений соответствующих переменных. Так, горизонтальная шкала охватывает промежуток от 100 до 350 единиц – количество высших учебных заведений, в то время как вертикальная шкала простирается от 500 до 800 – численность студентов. Нет необходимости говорить о том, что каждую ось следует разбить на достаточное число делений, чтобы обеспечить точное и сравнительно простое нанесение точек. В отличие от гистограммы, вертикальная шкала на поле не обязательно должна начинаться с нуля по той причине, что основное внимание здесь сосредоточено

на очертании рассеивания, а не на относительной частоте, которая оценивается по высоте кривой линии.

Начертив оси и прошкалировав их, можно изобразить каждую пару значений в виде точки на плоскости. Для каждой пары значений переменных, обозначенных точкой, значение *Y* определяет высоту точки над горизонтальной линией, а значение *X* определяет ее удаленность от вертикальной оси. Так, 1996–97 учебный год изображен на рисунке 4.3 точкой, расположенной на пересечении направляющих линий, восстановленных перпендикулярно к соответствующим осям из точек $Y = 595$ и $X = 274$. Для всех остальных пар чисел находятся «свои» точки в заданной системе координат, установленные на пересечении взаимно перпендикулярных прямых. Совокупность всех таких точек образует корреляционное поле.

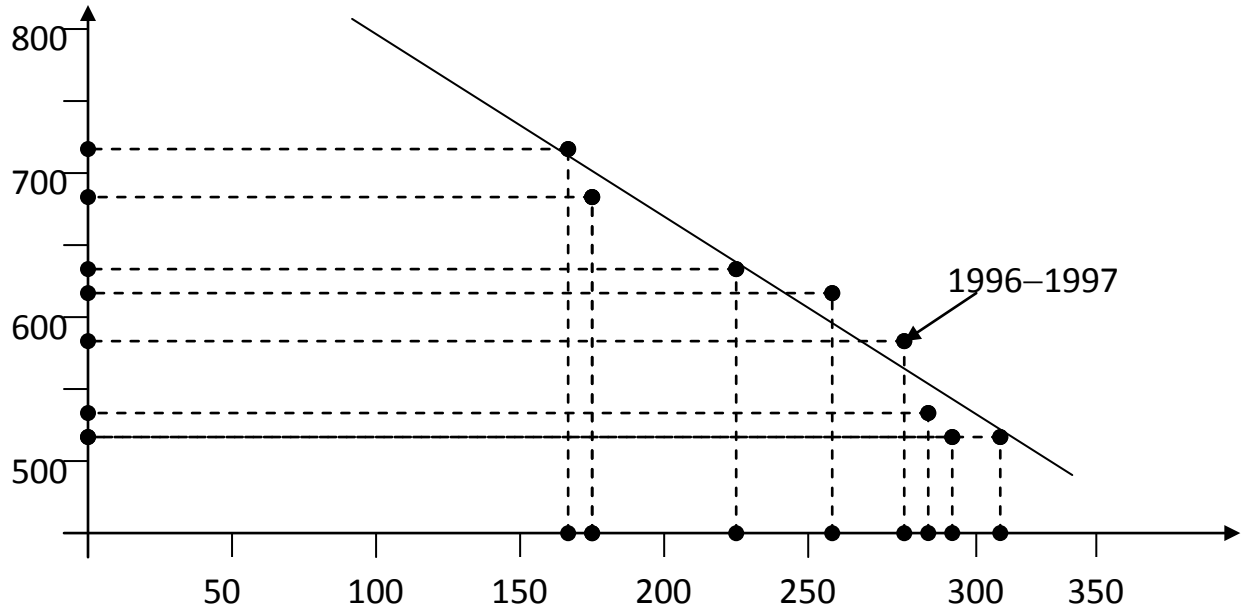

Рис. 4.3. Корреляционное поле динамики сети высших учебных заведений Украины III–IV уровня аккредитации и численности студентов в них

Виды рассеивания. Именно структура рассеивания позволяет судить о характере связи, а такое заключение является важной предпосылкой точного измерения связи. Так, например, как это явствует из анализа корреляционного поля, постепенное увеличение количества высших учебных заведений III–IV уровня аккредитации в Украине с 1992 по 2000 год сопровождается постепенным снижением численности студентов в них, то есть численность студентов изменяется на постоянную величину при изменении количества вузов на единицу измерения. Такая зависимость называется прямолинейной, или просто линейной, потому что общее направление распределения точек близко к прямой линии.

Любая такая линия концентрации данных, проведенная на глазок от руки или построенная на строгой математической основе, называется линией

регрессии. Это понятие было введено в обращение Ф. Гальтоном в 1877 году, который использовал его в связи с корреляционным исследованием некоторых особенностей родителей и их детей. Он заметил, что такая линия эффективно выражает тенденцию среди детей «регрессировать» к среднему уровню родителей по целому ряду признаков. Термин сохранился и получил широкое распространение, хотя его значение несколько изменилось.

Зависимость, изображенную на рисунке 4.4, называют обратной, так как два ряда значений переменных движутся в противоположных направлениях: по мере увеличения количества вузов численность студентов, обучающихся в них, падает. Если бы численность студентов и количество вузов возрастали одновременно, то общее направление рассеяния точек было бы снизу вверх и слева направо. Такая линия регрессии, как на рисунке 4.4, является доказательством прямой линейной зависимости между двумя переменными.

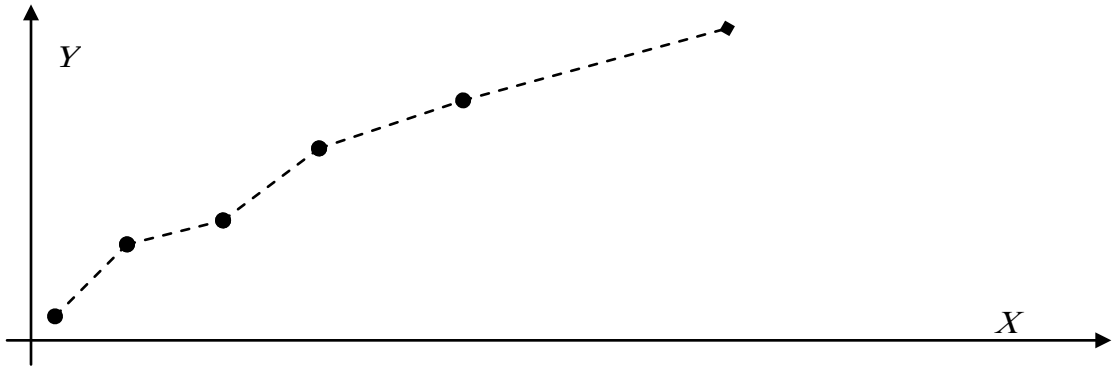

Рис. 4.4. Пример прямой ковариации

Тенденция рассеяния не всегда бывает линейной, чаще всего она бывает криволинейной и принимает форму одного из многочисленных видов кривых. На рисунке 4.5 дается пример, который изображает зависимость между благосклонностью отношения к национальному меньшинству и интенсивностью этого отношения. Решительное мнение – «За» или «Против» сочетается со значительной определенностью отношения и наоборот.

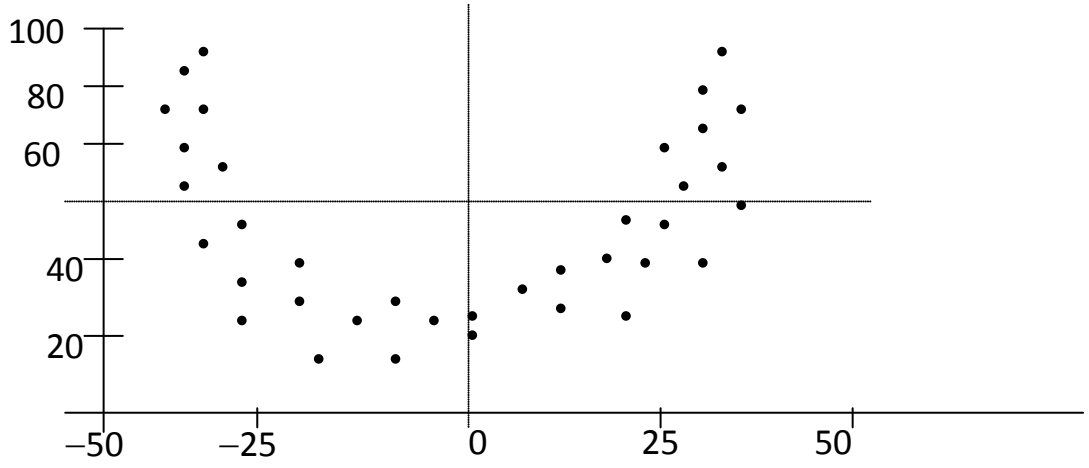

Рис. 4.5. Корреляционное поле (гипотетические данные)

Все три рассмотренных корреляционных поля имеют ярко выраженную тенденцию в рассеянии точек, поэтому каждый мог бы без особого труда охарактеризовать эти тенденции как линейные или криволинейные. Но распределения эмпирических наблюдений редко бывают столь определенными и недвусмысленными; более часто обе эти тенденции сочетаются в одной и той же совокупности данных и тем самым усложняют проблему выражения корреляции в виде обобщенной меры. Например, зависимость, представленная на рисунке 4.5, может быть интерпретирована как прямолинейная, однако, очевидно, что кривая линия больше соответствует тенденции рассеяния. По мере того как признак X увеличивается, признак Y также увеличивается, но в постепенно замедляющемся темпе. Это подтверждается выравниванием в скоплении точек. Наблюдение на корреляционном поле отклонений от общей кривой требует специального анализа, так как они представляют собой нарушение «закона» связи. Однако в основном закон связи выдерживается довольно хорошо, давая меру предсказуемости изменений одной переменной в зависимости от изменений другой. Если знаем, например, что количество высших учебных заведений III–IV уровня аккредитации Украины составляет 200, то можем предсказать, что численность студентов в них примерно равна 650 тыс., что соответствует высоте линии регрессии в этой точке (см. *Рис. 4.3*).

Несомненно, такое предсказание было бы не свободно от ошибки по той очевидной причине, что ни одно из наблюдаемых значений не попадает прямо на кривую в этой точке – все отклоняются в большей или меньшей степени. Очевидно, точность любого такого предсказания изменялась бы в зависимости от того, насколько плотно прилегают точки к линии, выражающей связь между двумя рядами данных. Точность предсказания, а следовательно, и корреляция, была бы высокой, если бы точки располагались концентрированно; когда точки рассеяны очень широко, точность предсказания бывает соответственно невысокой. Предсказание и корреляция были бы абсолютно полными только тогда, когда все точки располагались бы на линии регрессии. При другой крайности, когда рассеяние точек носит совершенно случайный характер, можно с равным успехом игнорировать так называемую «переменную-индикатор».

Скедастичность (вариабельность). Зная значение одного признака, не всегда можно с одинаковой вероятностью предсказать изменение другого, это связано со степенью рассеивания признака. Так, для примера, показывающего динамику сети высших учебных заведений Украины III–IV уровня аккредитации и численности студентов в них (см. *Табл. 4.2*), можно с одинаковой вероятностью предсказать изменение одного признака по изменению другого, так как рассеивание признака X – количества вузов равно 313–158=155, а рассеивание признака Y – численности студентов в этих вузах равно 718,8–503,7=215,1 тыс., что составляет числа одного порядка.

Рассеяние значений Y, соответствующих данному значению X, называется ***скедастичностью***. Если степень вариации значений (ширина зоны рассеяния) одинакова для всех значений X, то можно говорить о том, что переменная Y *гомоскедастична* по отношению к X. В противном случае, если, например, степень рассеяния значений Y уменьшается по мере изменения X, говорят, что переменная Y *гетероскедастична* по отношению к X. Гетероскедастичность означает, что степень корреляции неодинакова для всей совокупности данных, следовательно, ее наличие уменьшает возможность обойтись одним обобщенным показателем корреляции, который, в конце концов, всегда является средней величиной. Подобно тому, как не всегда можно вычислить среднее арифметическое для гетерогенных данных, точно так же избегают рассчитывать среднюю меру корреляции для гетероскедастического рассеяния.

Гетероскедастичность может быть ярко выраженной. Так, рассеяние может быть грушевидным, гантелеобразным или j-образным. Эти странного вида диаграммы рассеяния никоим образом не исчерпывают всех возможных типов распределения, которые могут встречаться в практической работе. Однако они все же пригодны для того, чтобы подтвердить полезность такого рода визуальных средств. Хотя диаграмма рассеяния не дает математической меры корреляции, она все же указывает: а) является ли зависимость простой прямолинейной или более сложной, б) является ли зависимость устойчивой для всех значений переменной, в) является ли связь сильной или слабой. Диаграмма рассеивания является необходимым средством анализа и играет при изучении ковариации ту же роль, что и график распределения частот при обработке одномерной совокупности данных. Она позволяет обозревать все распределение в целом.

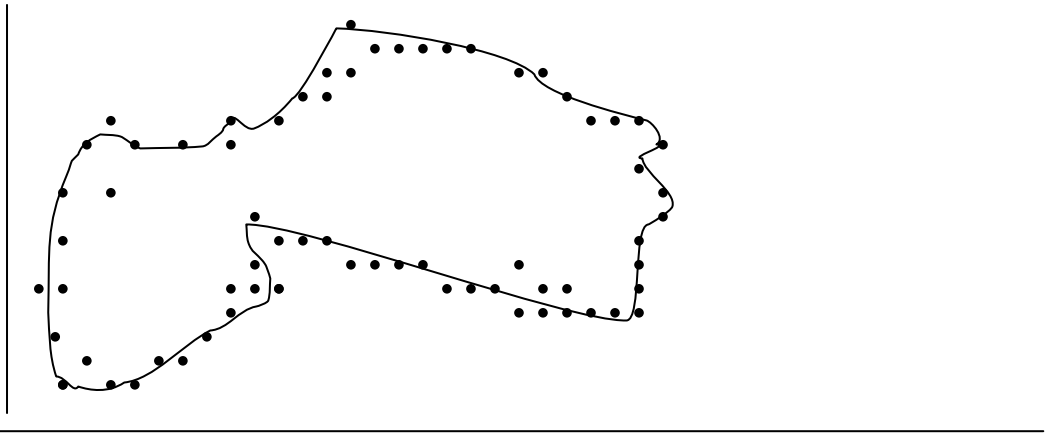

Рис.4.6. Диаграмма рассеяния

4.3. Корреляционная таблица

Вместо того чтобы наносить на чертеж отдельные наблюдения, можно сгруппировать их. Такая группировка отвечает следующим целям: 1)

определению основного вида рассеивания и 2) облегчению статистической обработки данных. При группировании двумерных совокупностей данные классифицируются так, что каждый случай учитывается одновременно по двуразрядным интервалам, тем самым располагая каждый случай в клетке на пересечении данной строки и данного столбца. Следовательно, получается так, как если бы наложили координатную сетку на корреляционное поле, подсчитали точки в клетках и вписали в каждую клетку соответствующее число. В результате применения такой операции и такого распределения получают таблицу распределения совместных частот (см. *Табл. 4.3*).

Естественно, в любой реальной ситуации не стали бы действовать таким несколько странным образом, скорее всего, непосредственно разнесли бы неупорядоченные двумерные наблюдения по клеткам координатной сети, специально разработанной для этой цели, и затем подсчитали бы количество наблюдений в каждой клетке.

Техника группирования. Чтобы уяснить технику группирования, перечислим правила, которых следует придерживаться:

1. Подбирают подходящий интервал группировки для каждой переменной.

2. Наносят эти интервалы на соответствующие оси координат и из каждой отметки проводят направляющие линии таким образом, чтобы получить координатную сетку. Предполагается, что каждая клетка представляет собой место пересечения двух интервалов группировки.

3. Помещают каждую связанную пару значений в соответствующую клетку и обозначают ее наличие каким-либо значком.

4. Подсчитывают значки в каждой клетке и заменяют их числом. Каждое такое число представляет собой совместную частоту, т. е. частоту, с которой появляются вместе значения или точки, относящиеся к двум признакам.

5. Суммируют по строкам и столбцам для того, чтобы получить маргинальные суммы. Они дают простые частотные распределения каждой переменной.

6. Суммируют маргинальные частоты, чтобы получить общее число всех случаев – N, причем, сумма частот по столбцам является контрольной цифрой для суммы частот по строкам и наоборот.

Таким образом, построение корреляционной таблицы происходит в соответствии с четко сформулированными принципами: классификация должна быть достаточно детальной, чтобы обнаружить форму распределения совместных численностей, и в то же время не настолько детальной, чтобы некоторые ряды и столбцы не остались совершенно пустыми. Но поскольку вначале неизбежны ошибки, то, вероятно, желательно иметь скорее избыток клеток в таблице, чем недостаток. Обычно всегда легче объединить клеточные частоты, чем разделить их.

Таблица 4.3

Корреляционная таблица совместных частот распределения времени (Y) в часах в неделю, которое тратят респонденты на занятия, не связанные с учебой в школе, в зависимости от класса обучения (X)

X\Y	до 1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	$\sum$
1	6										6
2	1	6									7
3	2	4	8	2		4					20
4			6	7	2	3					18
5			5	3	4	3					15
6				2	1	4	3				10
7						5	3	2	3		13
8							6	2			8
9								1	2	1	4
10								1	1	2	4
11							1	2	2	3	8
12									4	1	5
$\sum$ *(маргинальные сумы по столбцам)*	9	10	19	14	7	19	13	8	12	7	$N = 118$

Функция корреляционной таблицы. Необходимо признать, что корреляционная таблица не может изобразить вид зависимости так наглядно, как корреляционное поле (коррелограмма) по той причине, что цифры не могут передать столь же эффективно оттенки в плотности распределения, как это делает рассеяние точек. Тем не менее, несмотря на относительную грубость корреляционной таблицы, она часто оказывается вполне пригодной для установления вида зависимости и правильного ее понимания. Кроме того, распределения в строках и столбцах корреляционной таблицы позволяют произвести статистическое измерение скедастичности, что невозможно было бы сделать на основании исходного корреляционного поля. Для этого достаточно только вычислить квадратическое отклонение для каждого ряда или столбца. Близкое сходство между квадратическими отклонениями служило бы доказательством гомоскедастичности, тогда как заметные различия свидетельствовали бы о гетероскедастичности.

К тому же маргинальные распределения переменных X и Y, которые, безусловно, являются неотъемлемыми чертами корреляционной таблицы, также являются важным и общепризнанным средством для определения возможной

степени связи между переменными. В частности, маргинальные распределения устанавливают границы степени изменчивости получаемой корреляции. Например, непохожие маргинальные распределения исключают полную линейную зависимость. Обычно маргинальные распределения всегда действуют тем или иным образом в направлении ограничения силы связи между спаренными переменными; следовательно, всегда необходимо изучить их для того, чтобы определить ограничения, которые они накладывают, и возможности для обоснованного применения данного корреляционного показателя.

Корреляционная таблица может служить в качестве вспомогательного средства вычисления в тех случаях, когда число наблюдений велико или когда не имеется вычислительных машин. Наконец, она является предпосылкой для очень простого перехода к криволинейной корреляции, как будет показано в следующем параграфе.

В данном параграфе подчеркнута важная роль поля корреляции и корреляционной таблицы в предварительном анализе ковариации. С помощью этих диаграмм можно определить, является ли зависимость: а) линейной или криволинейной; б) прямой или обратной; в) слабой или сильной; г) являются ли переменные гомоскедастичными по отношению друг к другу; д) имеются ли значительные отклонения от основной тенденции или «закона связи» и е) являются ли маргинальные распределения симметричными или скошенными и сопоставимы ли они. Имея в виду получение этих важных предварительных сведений, всегда необходимо еще до измерения корреляции построить и тщательно изучить» диаграмму рассеяния или корреляционную таблицу, или даже и то и другое одновременно. Можно использовать неподходящие методы и, следовательно, прийти к неправильным выводам, если не прибегать к помощи этих визуальных средств.

Может это и покажется повторением, но учитывая то, что с началом следующего подраздела мы переходим к детальному рассмотрению коэффициентов связи, все-таки, напомним, что формулы измерения связи могут быть удобно сгруппированы на основе двух рассмотренных выше принципов связи: 1) *совместного появления* (коэффициент взаимной сопряженности Пирсона (C), стандартные коэффициенты связи Чупрова (T) и Крамера (T_c), коэффициент ассоциации Юла (Q) и коэффициент контингенции Пирсона (Φ); 2) *ковариации, взаимоизменяемости* (метод корреляции ранговых различий Спирмена (r_s) и Кенделла (τ), коэффициент корреляции Пирсона (r), коэффициент λ Гуттмана, корреляционное отношение (η), коэффициенты множественной ($R^2_{y(1\ldots k)}$) и частной корреляции ($r_{y1.2}$).

4.4. Коэффициенты связи, основанные на Хи-квадрат Пирсона (для номинальных признаков)

Номинальные признаки – это те признаки, которые измерены с помощью номинальных шкал. В первых параграфах данного издания мы говорили о том, что номинальные шкалы – шкалы низкого типа. А это значит, что возможное число измерительных процедур, которые можно осуществить применительно к таким шкалам, крайне ограничено. Считается, что для расширения исследовательских возможностей желательно обращаться именно к шкалам более высокого типа (порядковым, метрическим). Возможно, это, действительно так, но не в социологии. Согласимся с мнением тех социологов, которые утверждают, что роль номинальных данных в социологии огромна. В частности, Ю. Н. Толстова[17] объясняет это утверждение следующим образом [15, с. 164–165]:

- во-первых, именно номинальные данные чаще всего используются социологами, что объясняется сравнительной простотой их получения, естественностью интерпретации;
- во-вторых, номинальные данные являются более надежными, чем данные, полученные по шкалам более высокого типа, в том смысле, что за ними обычно не стоят трудно проверяемые модели восприятия (имеется в виду восприятие респондентом предлагаемых ему для оценки объектов, суждений, мнений и т. д.

Некоторые же авторы, например, С. В. Чесноков [18], вообще полагают, что в социологии только номинальные шкалы имеют право на существование [17]. Нередки случаи, когда количественные признаки, для которых более естественным кажется измерение с помощью метрических (числовых) шкал, вполне успешно измеряются с помощью номинальных шкал. Несмотря на то что номинальные шкалы являются шкалами низкого типа, для их анализа имеется немало эффективных методов. Наиболее часто социологи прибегают к расчетам коэффициентов, основанных на Хи-квадрат, то есть связанных с обязательным определением меры различия между наблюдаемыми (эмпирическими) и теоретическим частотами. Сам по себе Хи-квадрат является свидетельством связи между двумя признаками. Понятно, что при отсутствии связи величина Хи-квадрат равна нулю, и это значение является минимальным. В подразделе 3.4 данного издания мы приводили достаточно развернутый пример вычисления χ^2, в связи с чем не считаем необходимым обращаться к

[17] **Толстова Юлианна Николаевна**, д-р социол. наук, профессор кафедры сбора и анализа социологической информации Высшей школы экономики (г. Москва, Россия).

[18] **Сергей Валерианович Чесноков** (р. 29 июня 1943 г., СССР)российский ученый, математик, социолог, культуролог, музыкант, специалист по методам анализа данных и применению математических методов в гуманитарных исследованиях и проектах. Известен как создатель детерминационного анализа и детерминационной логики, исследователь гуманитарных оснований точных наук, активный участник песенного движения и артистического андеграунда в СССР и современной России.

подобным примерам вновь. Отметим, что основная проблема заключается в невозможности определения максимального значения χ^2, которое могло бы свидетельствовать о силе связи. Эта величина не имеет общего для всех таблиц сопряженности максимального значения, даже тогда, когда связь между признаками является максимально сильной (то есть когда каждому значению (категории) одного признака в точности соответствует значение другого признака). Кроме того, χ^2 зависит от числа степеней свободы, что делает невозможным сравнение между собой значений данной величины для таблиц с разным числом строк и столбцов. Эти аргументы, указывающие на несовершенство χ^2, приводят осознанию необходимости поиска коэффициентов, которые имели бы фиксированный максимум в случае максимальной связи и позволяли бы сравнивать между собой разные таблицы.

Среди коэффициентов, удовлетворяющих этим требованиям, социологами наиболее часто используются такие, как коэффициент сопряженности Пирсона (***C***), коэффициенты Чупрова (Т) и Крамера (T_C).

Коэффициент сопряженности Пирсона (*C*). Коэффициент сопряженности Пирсона (*C*) основывается на отклонении наблюдаемых частот в клетках таблицы от ожидаемых частот и предполагает, что распределение носит случайный характер [9, с.78−84]. Эти отклонения как раз и измеряются показателем χ^2, в соответствии с чем, формула вычисления коэффициента сопряженности Пирсона имеет следующий вид:

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

где ● χ^2 − вычисленная мера различия между эмпирическими и теоретическими частотами таблицы сопряженности признаков;

● *n* – число единиц наблюдения (объем выборки).

Эта формула, в которой учтено изменяющееся число наблюдений *n*, дает нормализованный показатель связи *C*. Необходимо обратить внимание на тот факт, что если χ^2 велико по сравнению с *n*, то C будет стремиться к единице, так как числитель и знаменатель фактически будут равны; однако, если χ^2 мало по сравнению с *n*, то коэффициент C будет также мал и в пределе будет стремиться к нулю. Если $\chi^2 = 0$ (то есть, если нет никакого расхождения между полученными данными и чисто случайным распределением), коэффициент C также будет равен нулю, потому что числитель равен нулю. Если все сказанное компактно представить в виде неравенства, будем иметь следующее: $0 \leq C < 1$.

Как видим, коэффициент сопряженности Пирсона *C* никогда не достигает единицы. Хотя и является очевидным, что чем ближе значение *C* к единице, –

тем сильнее связь, однако представляется необходимым нахождение $C_{\max}$ как максимального значения коэффициента, являющегося действительным эквивалентом полной корреляции.

$C_{\max}$ вычисляется по следующей формуле:

$$C_{\max} = \sqrt{\frac{\min(r-1; c-1)}{1+\min(r-1; c-1)}},$$

где
- r – число строк таблицы сопряженности;
- c – число столбцов таблицы сопряженности;
- *min* – означает, что принимается в расчет минимальная разность: либо от числа строк, либо числа столбцов.

Для квадратных таблиц уместно применить следующую формулу нахождения $C_{\max}$:

$$C_{\max} = \sqrt{\frac{t-1}{t}},$$

где t – число строк (либо столбцов – разницы нет).

Рассмотрим все сказанное на реальном примере. Повторимся, что алгоритм вычисления χ^2 описывался нами в третьем разделе данного издания (в подразделе 3.4), не будем повторяться, а просто подставим в формулу коэффициента сопряженности Пирсона C то значение χ^2, которое мы получили в том примере, пытаясь определить наличие/отсутствие связи между признаками. Однако, исходя из того, что в данном параграфе мы ведем речь только о номинальных признаках, предположим, что мы искали взаимосвязь между признаками «Профессиональные предпочтения» и «Социально-классовое происхождение», измеренными с помощью номинальных шкал. Итак, предположим, что мы получили $\chi^2 = 25{,}18$.

$$C = \sqrt{\frac{25{,}18}{25{,}18+154}} = 0{,}37 \approx 0{,}4,$$

$$C_{\max} = \sqrt{\frac{\min(4-1)(3-1)}{1+\min(4-1)(3-1)}} = \sqrt{\frac{\min 3; 2}{1+\min 3; 2}} = \sqrt{\frac{2}{1+2}} = \sqrt{0{,}7} = 0{,}8.$$

Вывод: учитывая значение $C_{\max}$, исходя из общих принципов интерпретации коэффициентов связи, можно сделать вывод о средней силе связи между признаками.

В заключение разговора о коэффициенте сопряженности Пирсона C еще

раз подчеркнем, что $C_{\max}$ является действительным эквивалентом полной корреляции, следовательно, этот показатель может быть использован в качестве стандарта для любой связи, которая меньше, чем полная (при условии, что распределения маргиналов являются идентичными). Отношение $C_{\max}$ и C является приблизительно эквивалентным условной мере связи, принимающей значения от «0» до «1». Чем больше размеры таблицы, тем ближе $C_{\max}$ к «1».

Коэффициенты сопряженности Чупрова (*T*) и Крамера (*Tc*). Уже сложившимся стандартом для измерения связи между признаками в прикладном социологическом исследовании являются еще два коэффициента сопряженности, тоже основанные на χ^2: это – коэффициенты Чупрова (*T*) и Крамера (*Tc*). Считается, что эти коэффициенты более строго оценивают тесноту связи, чем коэффициент сопряженности Пирсона (C).

Коэффициент Чупрова (T) находится с помощью следующей формулы:

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

где ● χ^2 – вычисленная мера различия между эмпирическими и теоретическими частотами таблицы сопряженности признаков;

● *n* – число единиц наблюдения (объем выборки);

● *r* – число строк таблицы сопряженности;

● *c* – число столбцов таблицы сопряженности.

Значение, принимаемое коэффициентом Чупрова, может варьироваться от «0» до «1», соответствующее неравенство имеет следующий вид: $0 \le T \le 1$.

Как видим, значение данного коэффициента может быть равным единице, однако только в случае квадратной матрицы (то есть такой, которая имеет одинаковое число строк и столбцов; r = s).

В том случае если $r \neq s$, необходимо, как и в случае с коэффициентом сопряженности Пирсона (*C*), найти действительный эквивалент полной корреляции или T_{max}. Формула вычисления этого значения имеет следующий вид:

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

где ● *r* – число строк таблицы сопряженности;

● *c* – число столбцов таблицы сопряженности;

● *min* – означает, что принимается в расчет минимальная разность: либо от числа строк, либо от числа столбцов;

● *max* – означает, что принимается в расчет максимальная разность: либо от числа строк, либо от числа столбцов.

Проиллюстрируем сказанное на примере, который рассматривался выше. С помощью коэффициента Чупрова (Т) определим силу связи между признаками «Профессиональные предпочтения» и «Социально-классовое происхождение», для которых $\chi^2 = 25{,}18$.

$$T = \sqrt{\frac{25{,}18}{154\sqrt{(4-1)\times(3-1)}}} = \sqrt{\frac{25{,}18}{377{,}3}} = \sqrt{0{,}067} = 0{,}26 \approx 0{,}3;$$

$$T_{\max} = \sqrt[4]{\frac{\min(4-1;3-1)}{\max(4-1;3-1)}} = \sqrt[4]{\frac{2}{3}} = \sqrt[4]{0{,}67} \approx 0{,}9.$$

Вывод: учитывая значение $T_{\max}$, очень приближенное к единице, исходя из общих принципов интерпретации коэффициентов связи, можно сделать вывод об умеренной (однако, ближе к слабой) силе связи между признаками.

Возникает вопрос, в пользу какой связи – слабой или умеренной – сделать окончательный вывод? И без того спорная ситуация, связанная с тем, что предварительно полученный коэффициент сопряженности Пирсона свидетельствует в пользу более сильной связи между признаками, усложняется еще и проблемой округления полученного значения. Дело в том, что значение 0, 26 – это еще не умеренная, а слабая связь, однако если его округлить до 0,3 (что имеем полное право сделать), можно получить показатель, свидетельствующий об умеренной связи. Как поступить в такой ситуации? Каков должен быть окончательный вывод? Тут-то как раз и «приходит на помощь» коэффициент Крамера. И неспроста эти два коэффициента, Чупрова и Крамера, обычно упоминаются в паре. Значения этих коэффициентов как бы «подкрепляют» друг друга.

Коэффициент Крамера (Тс)[19] также является мерой связи двух номинальных переменных, основанной на критерии X^2 . Однако, учитывая то, что за основу вычисления, как правило, берется значение коэффициента Чупрова, можно сказать, что ***Тс*** является уточняющим по отношению к ***Т***. Для нахождения коэффициента Крамера применяется следующая формула:

$$Tc = \frac{T}{Tmax},$$

где
- T – значение коэффициента Чупрова;
- T_{max} − максимально возможное значение коэффициента, являющееся действительным эквивалентом полной корреляции.

Значение, принимаемое коэффициентом Крамера, также может варьироваться от «0» до «1», соответствующее неравенство имеет следующий вид: $0 \le T_c \le 1$.

[19] В некоторых источниках можно встретить использование «V» в качестве условного обозначения для коэффициента Крамера.

Рассмотрим сказанное на примере и найдем коэффициент Крамера для измерения силы связи между теми же признаками «Профессиональные предпочтения» и «Социально-классовое происхождение», для которых $\chi^2 = 25{,}18$. В итоге получим:

$$Tc = \frac{0{,}3}{0{,}9} = 0{,}33\,.$$

Вывод: исходя из общих принципов интерпретации коэффициентов связи, можно сделать вывод об умеренной силе связи между рассматриваемыми признаками.

Как видим, приведенный выше пример как нельзя лучше объясняет, почему одного коэффициента недостаточно для измерения связи между признаками.

4.5. Коэффициенты связи для матриц *2×2* (для номинальных и порядковых признаков)

Для определения статистической связи переменных, измеренных дихотомической шкалой наименований (то есть при помощи номинальной шкалы, содержащей только две альтернативы), используются коэффициенты *ассоциации* (Юла, Q) и *контингенции* (Фи, Φ).

Коэффициент ассоциации Юла (Q). Чтобы иметь единую меру связи для матриц ***2×2***, английский статистик Д. Юл предложил следующий коэффициент связи, который обозначил «Q» в честь известнейшего ученого XIX века А. Кьютелета (Quetelet)[20]. Соответствующая формула имеет следующий вид:

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

где a, в, c, d являются частотами, распределенными в четырехклеточной таблице (матрице **2×2**) следующим образом:

Признаки	***A***	***He-A***
B	a	b
He-B	c	d

Значение, принимаемое коэффициентом ассоциации, может варьироваться от «−1» (при полной обратной связи) до «1» (при полной прямой связи) и быть равным «0» при отсутствии статистической зависимости, а

[20] **Lambert Adolphe Jacques Quetelet** (22 февр. 1796 – 17 февр. 1874) – бельгийский астроном, математик, статистик, социолог, прославившийся успешным использованием математико-статистических методов в социальных науках.

значит, и связи между признаками. Соответствующее неравенство имеет следующий вид: $-1 \le Q \le 1$.

При этом, чем ближе значение коэффициента Q к нулю, − тем слабее связь между признаками.

Применим приведенную выше формулу вычисления коэффициента ассоциации Q при обработке данных таблицы 4.4.

Таблица 4.4

Распределение правонарушителей по полу

	Юноши	*Девушки*	Итого
Правонарушители	20	0	20
Не-правонарушители	50	50	80
Итого	50	50	100

Как видим из таблицы, в гипотетической группе из 100 человек имеется 20 правонарушителей, и все они – юноши. В данном случае возможно точное предсказание, то есть существует полная определенность относительно пола правонарушителя. Исходя из имеющихся данных, можно сказать, что преступность всецело объясняется признаком пола, так как ни одна из девушек не входит в группу правонарушителей. Однако в социологической практике чаще встречаются не такие очевидные ситуации, что требует дополнительных вычислительных процедур для измерения точной корреляции между признаками. Осуществим такую процедуру с использованием соответствующей формулы для вычисления коэффициента ассоциации Юла (Q) и получим:

$$Q = \frac{20 \cdot 50 - 30 \cdot 0}{20 \cdot 50 + 30 \cdot 0} = \frac{1000 - 0}{1000 + 1} = 1.$$

Такое значение показателя (Q=1) является убедительной мерой полной связи между полом и склонностью к совершению преступления. Очевидно, что коэффициент ассоциации достигает единицы в ситуации, когда *в* = 0 либо *с* = 0. Следовательно, сам коэффициент ассоциации показывает нам не только силу и направление связи. Зная этот коэффициент, мы можем сделать конкретный вывод и относительно характера этой связи, то есть сказать с уверенностью, что именно юноши, а не девушки являются правонарушителями. Если сказать более строго, то коэффициент ассоциации измеряет одностороннюю связь, в то время как коэффициент контингенции (Фи), о котором речь пойдет далее, измеряет двустороннюю связь между признаками и на его основе мы не смогли бы сделать вывод относительно характера этой связи. То есть зная о наличии сильной прямой связи между полом и склонностью к правонарушениям, конкретизировать, каков же пол правонарушителей.

Коэффициент контингенции Фи (Φ). Надо сказать, что статистические свойства Φ аналогичны Q, но в некоторых отношениях они отличаются друг

от друга. Подобно Q, этот показатель применим только к дихотомическим таблицам ***2×2***, не фиксирующим градаций в значениях признаков, или к непрерывным переменным, которые могут быть более или менее обоснованно дихотомизированы. Точно так же как и Q, Φ измеряет интенсивность связи, выражаемой распределением частот в клетках таблицы. Что касается формул для Φ и Q, то числители у них идентичны, а знаменатели сконструированы различным образом. Формула вычисления коэффициента контингенции *Фи* имеет следующий вид:

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

где а, в, с и *d* – являются уже известными частотами четырехклеточной таблицы, матрицы ***2×2,*** а суммы отдельных клеточных частот являются маргиналами.

Значение, принимаемое коэффициентом контингенции, может варьироваться от «−1» (при полной обратной связи) до «1» (при полной прямой связи) и быть равным «0» при отсутствии статистической зависимости, а значит и связи между признаками. Соответствующее неравенство имеет следующий вид: $-1 \le \Phi \le 1$.

При этом, чем ближе значение коэффициента Φ к нулю, тем слабее связь между признаками. Однако, если коэффициент ассоциации Юла достигает единицы в ситуации, когда b = 0 либо c = 0, то коэффициент контингенции Фи достигает единицы в иной ситуации, а именно, когда a = d = 0 или b = c = 0.

Для того чтобы продемонстрировать различия в процессе применение Q и Φ, можно обратиться опять к перекрестной классификации правонарушителей по полу в случае, когда Q был равен единице (см. *Табл. 4.4*). Подставляя в формулу для Φ те же самые данные, получим значение коэффициента в два раза меньше, а именно:

$$\Phi = \frac{1000 - 0}{\sqrt{20 \cdot 80 \cdot 50 \cdot 50}} = \frac{1000}{2000} = 0{,}5.$$

Помним, что в этом случае Q равен единице, так как все правонарушители являются мальчиками. Однако обратное утверждение, что все мальчики являются правонарушителями, очевидно, было бы неверно. Отсюда коэффициент Φ, формула которого приспособлена для того, чтобы отразить эту взаимную зависимость, равен всего лишь 0,5. Связь является взаимной, но не идеальной (абсолютной).

Несомненно, коэффициенты Q и Φ измеряют различные аспекты взаимосвязи в четырехклеточной таблице. По существу, различие между двумя

показателями заключено в том факте, что формула для Φ позволяет отразить степень двусторонней взаимосвязи с помощью единичного индекса. Коэффициент Φ «схватывает» любую двустороннюю взаимосвязь, которая существует между двумя рядами значений качественных признаков. Так как данный показатель измеряет только двустороннее отношение между X и Y, то он в равной мере учитывает воздействие обеих переменных, поэтому его можно назвать обратимым.

Это можно очень просто проиллюстрировать (см. *Табл. 4.5*) в случае совершенной двусторонней взаимосвязи − ситуации, в которой нет отклонений ни в ту, ни в другую сторону: все мальчики являются правонарушителями, а все правонарушители − мальчиками.

Таблица 4.5

Пример идеальной двусторонней связи между полом и правонарушениями

	Мальчики	*Девочки*	Итого
Правонарушители	20	0	20
Не-правонарушители	0	80	80
Итого	20	80	100

В данном случае коэффициент контингенции также равнялся бы единице:

$$\Phi = \frac{1600 - 0}{\sqrt{20 \cdot 80 \cdot 20 \cdot 80}} = \frac{1600}{1600} = 1.$$

Вследствие этой совершенной двусторонней взаимосвязи, наблюдения распределяются вдоль одной диагонали матрицы ***2×2***, а клетки, расположенные на другой диагонали, остаются пустыми. Поскольку каждый признак всецело «объясняет» другой, то Φ должен быть равен единице. Если выразить это в арифметической форме, то два нуля на одной из диагоналей таблицы с необходимостью приводят к Φ=1. Этот принцип обратимости также хорошо выдерживается для любого промежуточного значения от нуля до единицы, то есть когда спаренные признаки только частично объясняют друг друга. По мере того как степень взаимной зависимости снижается, значение Φ также уменьшается.

Коэффициенты Q и Φ, конечно, совпадают при наличии идеальной двусторонней взаимосвязи по той простой причине, что полная односторонняя связь (Q=1) является необходимым элементом совершенной двусторонней связи (Φ=1).

4.6. Коэффициенты связи, основанные на моделях прогноза (для номинальных признаков)

Что имеется в виду, когда говорится о прогнозных математических моделях? Ответ прост: в данном случае речь идет о взаимосвязанных (коррелирующих) признаках, однако связь эта такова, что наблюдение за «поведением» одного признака дает возможность предвидеть «поведение» другого. Или другими словами: признаки будут считаться взаимосвязанными, если реализованное значение одного из них позволяет относительно точно предугадать, каким будет значение другого [3, с. 143−180; 7, с. 327].

Например, если мы хотим опросить людей предпенсионного возраста на предмет их отношения к новой пенсионной реформе, нам и в голову не придет искать респондентов в ночном клубе или диско-баре, так как мы точно знаем, что средний возраст посетителей подобных увеселительных заведений не превышает тридцати лет. Получается, что признак «1»− «место нахождения» находится в тесной взаимосвязи с признаком «2» − «возраст». Более того, не зная значений признака «2» (конкретных возрастных показателей), мы можем с большой степенью вероятности их угадать, зная значения признака «1» (конкретные заведения).

Однако в данном случае правомерность сделанного нами вывода подтверждается, скорее, обыденным знанием, чем научным. Тем не менее есть вполне научные методы, с помощью которых, используя соответствующие формулы и осуществив необходимые вычисления, можно спрогнозировать ситуацию, обосновав этот прогноз с помощью конкретных (относительно точных) цифр и показателей.

Для измерения такого рода корреляционной зависимости могут быть применены меры λ (лямбда) Гуттмана. Следует отметить, что в специальной литературе можно встретить формулы вычисления коэффициента λ Гуттмана, записанные как через абсолютные частоты, так и через частости. В данном учебнике мы решили не приводить их все, чтобы не утяжелять текст излишними формулами, что, на наш взгляд, усложнит восприятие и усвоение самой важной информации.

Кроме того, важно помнить, что фамилия Гуттмана в разных учебниках, учебных пособиях и другой специальной научной литературе приводится по-разному, например: Гу*т*м*е*н, Гу*т*ман Гу*д*ман. Такие расхождения связаны с «трудностями» перевода. Надо понимать, что если речь идет о коэффициенте корреляции и при этом упоминается какой-либо вариант фамилии из списка, представленного выше, − на самом деле имеется в виду один и тот же человек, известный всему миру социолог, статист, психолог Луи Гуттман[21].

[21] **Louis (Eliyahu) Guttman** *(Бруклин,1916−Минеаполис,1987)* известный ученый, прославившийся своей изощренностью в применении математических методов в социологии и психологии. Блестящий новатор, который «видел теорию в методе и метод теории». Один из самых влиятельных представителей психометрического анализа 20-го столетия. В 1942 получил Степень Ph.D. (доктора философии) по социологии по теме, связанной с факторным анализом (Университет

Вообще, существует три меры λ Гуттмана: две из них− направленные, а одна представляет собой усреднение первых двух. Направленная мера показывает силу зависимости поведения одного признака от поведения другого. Например, направленная мера Гуттмана λ_{yx} («лямбда_игрик_по_икс») показывает, как изменяются значения признака *Y* в зависимости от изменения значений признака *X*, то есть, по сути, силу влияния *X* на *Y*. Формула направленной меры (коэффициента) Гуттмана λ_{yx} выглядит так:

$$\lambda_{yx} = \frac{\sum \max n_{ij} - \max n_{ci}}{n - \max n_{ci}},$$

где
- $\max n_{ij}$ − максимальные частоты по строкам;
- $\max n_{ci}$ − максимальная маргинальная частота по столбцам;
- n − объем выборочной совокупности.

Алгоритм вычисления второй направленной меры Гуттмана λ_{xy} («лямбда_икс_по_игрик») аналогичен приведенному выше, просто меняются местами столбцы и строки таблицы сопряженности, а сам коэффициент показывает, как изменяются значения признака *X*, в зависимости от изменения значений признака *Y*.

Третья, усредненная, мера показывает общую силу связи между признаками и вычисляется довольно просто − путем нахождения среднего арифметического из двух направленных мер. Соответствующая формула вычисления коэффициента λ Гуттмана имеет следующий вид:

$$\lambda = \frac{\lambda_{yx} + \lambda_{xy}}{2}.$$

Теперь, когда формулы известны, рассмотрим вычисление всех трех мер (коэффициентов) λ Гуттмана на конкретном примере и попытаемся сделать научно обоснованный прогноз.

Представим себе гипотетическую ситуацию измерения связи между профессией человека и уровнем удовлетворенности жизнью. Попытаемся определить, сможем ли мы сделать прогноз о том, насколько человек удовлетворен своей жизнью, зная только его профессию (λ_{xy}). Или наоборот, сможем ли мы отгадать профессию человека, зная только то, насколько он

Миннесоты). В 1947 иммигрировал в Израиль, где основал и возглавил израильский Институт прикладного социального исследования (позже – Институт Гуттмана). С 1955 профессор Социальной и Психологической Оценки в Еврейском Университете Иерусалима, где работал практически до последних дней жизни. Авторству Гуттмана принадлежит множество публикаций и научных книг по социологии, психологии и статистике, многие из которых получили всемирную известность.

удовлетворен своей жизнью. В таблице 4.6 представлены абсолютные частоты двумерного распределения по обозначенным признакам. Для облегчения вычислительных процедур в эту таблицу мы добавили дополнительный столбец (с максимальными частотами по строкам) и дополнительную строку (с максимальными частотами по столбцам).

Таблица 4.6

Результаты двумерного распределения по признакам «Профессия» и «Удовлетворенность жизнью»

Профессия (X)	*Количество респондентов по уровню удовлетворенности жизнью (Y)*			*Максимальная частота*	*Общее число респондентов*
	1	***2***	***3***		
Учитель	5	8	*18*	*18*	31
Продавец	12	7	*14*	*14*	33
Повар	7	11	*20*	*20*	38
Водитель	*19*	10	6	*19*	35
Максимальная частота	*19*	*11*	*20*		***50***
Итого	43	36	58	***71***	137 (n)

Подставив данные этой таблицы в соответствующие формулы получим:

$$\lambda_{yx} = \frac{(18+14+20+19) - \max 43;36;\mathbf{58}}{137 - \max 43;36;\mathbf{58}} = \frac{71-58}{137-58} \approx 0{,}16.$$

$$\lambda_{xy} = \frac{(19+11+20) - \max 31;33;38;35}{137 - \max 31;33;38;35} = \frac{50-38}{137-38} \approx 0{,}12.$$

$$\lambda \approx \frac{0{,}16+0{,}12}{2} \approx \frac{0{,}28}{2} \approx 0{,}14.$$

Результаты вычислений показывают, что в целом связь между признаками довольно слабая. Это касается всех трех мер λ Гуттмана. Даже по тому, как вычисляется коэффициент, видно, что он позволяет определять, существуют ли в строках модальные группы, то есть есть ли в каждой профессиональной группе ярко выраженная, часто встречаемая «степень удовлетворенности жизнью». Судя по нашей таблице, таких групп практически нет, что и подтверждается маленьким значением коэффициента.

Какими же свойствами обладает этот коэффициент? Его значения могут варьироваться от «0» до «1», соответствующее неравенство имеет следующий вид: $0 \leq \lambda \leq 1$.

В случае, когда значение коэффициента равно «1», вероятность статистического предсказания У по Х максимальная. Такой случай практически в социологических исследованиях не встречается. Если рассмотреть его на нашем примере, то равным единице коэффициент λ Гуттмана мог бы быть только в том случае, если бы в каждой профессиональной группе все респонденты имели одинаковую степень удовлетворенности жизнью, и при этом в каждой из групп эта степень удовлетворенности была бы «своей».

Значение λ Гуттмана, равное нулю, свидетельствует о том, что знание признака Х нечего не даст нам для увеличения знания о признаке У и/или наоборот.

Представляется важным отметить, что в реальных исследованиях значения коэффициента Гуттмана очень малы и использовать их нужно, как и многие другие коэффициенты, в сравнительном контексте. Например, для ранжирования «как бы» независимых между собой признаков по степени их влияния на какой-то конкретный, особенно важный для исследователя, признак (обозначаемый как целевой или зависимый).

4.7. Корреляция рангов

Измерение с помощью рангов. Переменные могут быть связанными таким образом, что вариация одной переменной соответствует вариации другой и тогда о них говорят, что они ковариантные. Уровень рождаемости и уровень дохода в семье, число самоубийц и доля верующих – все это может быть связано друг с другом, а степень связи измерена с помощью подходящего показателя корреляции. Из всего множества известных корреляционных индексов мы обратимся к рассмотрению только двух – коэффициента парной ранговой корреляции Спирмена (r_s) и коэффициента конкордации (W), который определяет степень множественной связи между качественными признаками.

Коэффициент ранговой корреляции используется для измерения взаимозависимостей между качественными признаками, значения которых могут быть упорядочены или проранжированы по степени убывания (или нарастания) данного качества у исследуемых социальных объектов.

Самый простой способ упорядочения данных состоит в их ранжировании. Простота этого способа заключается в том, что измерение может производиться интуитивно и субъективно, а не посредством явных, объективных единиц измерения. Такое упорядочение не может быть очень точным, и все же оно иногда оказывается очень полезным. В сущности, многие социологические понятия и не могут быть упорядочены никаким иным образом. Значительное число социологических исследований базируется на подобной субъективной основе. Так, профессии могут быть ранжированы по престижу; коллектив студентов может быть ранжирован в порядке предпочтения; национальности могут быть расположены в ряд посредством благосклонного или неблагосклонного к ним отношения; художественные картины могут быть ранжированы по эстетической ценности. Даже когда единицы измерения существуют как, например, при определении размеров городов или величины

рождаемости, к ранжированию можно прибегнуть в том случае, если подобная точность не требуется. Так как ранговый порядок основывается на принципе «больше или меньше», можно рассматривать совокупность упорядоченных наблюдений как количественную, а не качественную переменную. Когда два ряда рангов изменяются совместно, то можно говорить о ранговой корреляции.

Коэффициент ранговой корреляции Спирмена (r_s). Простейший случай корреляции рангов – корреляция только между двумя рядами рангов. Когда число ранжированных объектов не превышает шести, наблюдатель в состоянии путем простого обозрения вывести грубое, но приемлемое заключение о степени связи между ними. Рассмотрим два ряда расположения шести картин двумя экспертами, как показано в таблице 4.7.

Таблица 4.7

Пример полной положительной корреляции

Картина	Эксперт X	Эксперт Y	Различия (*d*)
A	1.	1.	0
B	2.	2.	0
C	3.	3.	0
D	4.	4.	0
E	5.	5.	0
F	6.	6.	0

Поскольку эти ряды дублируют друг друга, то различие между каждой парой рангов равно нулю. Кроме того, имеется полное соответствие одного рангового порядка другому. Таким образом, мерой корреляции между рангами должна быть единица. С другой стороны, ряд рангов можно было бы упорядочить от низшего уровня к высшему, никак не повлияв на степень предсказуемости, вместо того, чтобы сделать это упорядочение от высшего уровня к низшему (см. *Табл. 4.8*).

Таблица 4.8

Пример полной отрицательной корреляции

Картина	Эксперт X	Эксперт Y	Различия (*d*)	d^2
A	1.	6	–5	25
B	2.	5.	–3	9
C	3.	4.	1	1
D	4.	3.	1	1
E	5.	2.	3	9
F	6.	1.	5	25
Итого			0	70

Несмотря на то, что два ряда рангов находятся сейчас по отношению друг к другу в строго обратном порядке, отклонение каждого ранга от среднего ранга остается неизменным, и поэтому взаимная предсказуемость их значений

остается той же самой. Отсюда мерой связи все еще является единица, но уже с отрицательным знаком.

Третий возможный вид зависимости – между этими двумя крайностями – случайная связь, при которой ранг X нельзя предсказать по рангу Y (см. *Табл. 4.9*); любой X–ранг с равной степенью вероятности мог бы сочетаться с любым Y–рангом. В этом случае, по крайней мере, теоретически, корреляция была бы равна нулю.

Таблица 4.9

Пример случайной связи

Картина	Эксперт X	Эксперт Y	Различия (d)	d^2
A	1.	3	–2	4
B	2.	5.	–3	9
C	3.	1.	2	4
D	4.	6.	–2	4
E	5.	2.	3	9
F	6.	4.	2	4
Итого			0	34

Формула для измерения степени корреляции между двумя рядами рангов, приспособленная к условному диапазону колеблемости значений коэффициента от 0 до 1, была выведена Спирменом в 1904 году и имеет следующий вид:

$$r_s = 1 - \frac{6 \times \sum d^2}{N \times (N^2 - 1)},$$

где
- d – разность между парными рангами,
- N – число ранжированных объектов.

Решая три вышеприведенных примера, получим следующие коэффициенты:

Таблица 4.7

$$r_s = 1 - \frac{6 \cdot 0}{6(36-1)} = 1$$

Таблица 4.8

$$r_s = 1 - \frac{6 \cdot 70}{6(35)} = -1$$

Таблица 4.9

$$r_s = 1 - \frac{6 \cdot 34}{6(35)} = 0{,}03$$

Как видно, величина r_s может изменяться в пределах от −1 до +1, когда два ряда проранжированы в одном порядке. Соответствующее неравенство имеет следующий вид: $-1 \le r_s \le 1$.

При полном взаимном беспорядочном расположении рангов $r_s = 0$.

Значимость коэффициента корреляции Спирмена можно определить по таблицам критических величин r_s (см. Приложения, табл. «А», «Г»). Наблюдаемые значения критерия вычисляются по следующей формуле:

$$z = \frac{r_s}{\sqrt{N-1}}.$$

Следует отметить, что социологи наиболее часто прибегают к вычислению коэффициента корреляции Спирмена r_s в тех случаях, когда необходимо сравнить две различные выборки. Например, когда мы хотим сравнить, насколько отличаются/совпадают ценностные ориентации студентов первых курсов в двух различных вузах или жизненные планы старшеклассников нынешнего и прошлого года (однако при условии, что соответствующие измерения по обозначенным признакам на разных выборках осуществлялись с помощью одних и тех же шкал).

Техника вычисления в случае объединенных рангов. Ранги иногда могут объединяться. Эксперт в затруднительном положении может оценить две картины одинаково, количественные значения и меры могут также быть одинаковыми. В подобных случаях два или более наблюдений могут, по-видимому, одновременно претендовать на один и тот же ранг. Так как число рангов и число наблюдений должно совпадать, то просто невозможно двум наблюдениям присвоить один и тот же ранг, поэтому им должны быть присвоены объединенные ранги. В таких случаях обоим наблюдениям приписывается значение среднего арифметического из двух объединенных рангов. Так, если объединяются 3-е и 4-е наблюдения, каждому присваивается ранг 3,5. Если объединяются три (и более) ранга, действует то же самое правило усреднения. Поскольку такие объединения затушевывают различия рангов, они приводят к уменьшению предельного значения корреляции.

Корреляция между упорядоченными переменными. Если даны относительно короткие ряды количественных данных и если нет необходимости в свойственной им точности, можно расположить эти данные в порядке их величин и коррелировать ранги вместо числовых значений. Преимущества этого приема заключаются в быстроте и простоте вычисления, что компенсирует уменьшение точности корреляционного показателя. В таблице 4.10 коррелируются доход и месячная квартплата шести семей посредством сравнения ранговых порядков.

Таблица 4.10

Ранжирование семей по доходу и месячной квартплате

Семья	*Доход*	*Кв. плата*	*Ранг*
А	1500	250	1
Б	700	200	2
В	500	125	3
Г	450	90	4
Д	400	80	5
Е	300	75	6

Как видно из таблицы, столбцы значений находятся в полном ранговом соответствии: ранговая градация по доходу точно соответствует градации по квартплате. Однако исходные величины не обнаруживают такого же полного соответствия. Конечно, такие ряды спаренных рангов дадут r_s, равное единице, но только потому, что игнорируют детали исходной информации. Практически, для доказательства того факта, что корреляция между количественными данными не равна 1, будет вполне достаточно для построения корреляционного поля, поскольку некоторые точки на нем будут отклоняться от прямой линии, которая изображает полную линейную корреляцию.

Анализ $r_s \times r_s$ можно действительно рассматривать в качестве меры конгруэнтности (или согласия), колеблющейся от полного совпадения (+1) через случайную связь (0) к полному несоответствию (–1) между рядами рангов. Отсюда очевидна обоснованность базирования формулы на различиях между спаренными рангами.

Однако различия между рангами имеют более точный статистический смысл, который не виден из формулы. Но студент должен уже сознавать, что формулы часто являются в высшей степени конденсированными, операциональными конструкциями, которые скрывают больше, чем обнаруживают. Поэтому совсем не очевидно, что r_s коррелирует скорее с сигма-значениями (σ), а не с рангами. Уже имели возможность заметить, что ранговая корреляция основывается не на прямом соответствии между ранговыми порядками, а скорее на соответствии между отклонениями соответствующих рангов от среднего ранга. Эти отклонения измеряются в сигма-единицах. Формула автоматически превращает ранги в σ и отсюда действительные различия между рангами (d) − в различия, выраженные в сигмах (σ). Следовательно, можем подставлять в нее исходные данные, упорядоченные по рангам.

Однако вышеупомянутая серия операций основывается на предположении, что интервалы между рангами равны. Например, предполагается, что разрыв между рангами 1 и 2 равен разрыву, который существует между рангами 2 и 3. И все же здравый смысл подсказывает, что идеального случая равных интервалов, которые требуются для формулы, достигнуть невозможно.

Действительная степень субъективного предпочтения первого ранга по сравнению со вторым не является обязательно такой же по интенсивности, как предпочтение второго ранга по сравнению с третьим. Так, лошадей можно расположить в порядке пересечения ими финишной линии, но интервалы между ними не будут одинаковыми. Тем не менее, r_s может использоваться и используется в том случае, когда отсутствуют объективные единицы измерения

или когда различие между неравными интервалами рассматривается как несущественное.

Из вышесказанного становится понятным, что r_s измеряет корреляцию между порядковыми рангами, а не между ранжируемыми величинами. Следовательно, r_s преувеличивает степень связи между изучаемыми переменными. Таким образом, два эксперта, упорядочивающие одну и ту же совокупность художественных полотен, могут иметь различный художественный вкус и все же расположить произведения искусства в идентичном порядке. Эксперт 1 может считать все их, в высшей степени, превосходными, тогда как эксперт 2 будет оценивать одинаково низко всю совокупность картин. Эксперты могут согласиться относительно порядка, но разойтись во мнении по существу. Несмотря на все это, для большинства реальных ситуаций разумно предположить, что ввиду общности культуры сходное ранжирование всегда соответствует сходным предпочтениям. Когда два ряда данных не располагаются на единой непрерывной шкале, как в случае арендной платы и дохода, или когда они относятся к несоизмеримым категориям, как, например, рождаемость и доход, то подобная проблема обоснования их расположения на шкале, естественно, не возникает. И все же, прежде чем использовать r_s, исследователь должен убедиться в том, что его интересует корреляция только между рангами, а не между фактическими величинами.

Мера соответствия для трех и более ранговых рядов. Формула для вычисления r_s пригодна только для двух ранговых рядов, однако данные могут состоять из трех и более ранговых рядов. Один из методов определения общей степени соответствия между тремя и более упорядоченными рядами данных состоит просто в вычислении среднего арифметического из всех возможных значений r_s. Таким образом, если бы три эксперта ранжировали шесть картин, можно было бы вычислить r_s для всех возможных сочетаний рядов по парам для того, чтобы определить среднее согласие между ними. Спаренные ряды сочетались бы следующим образом: эксперты 1 и 2, 2 и 3, 1 и 3. Затем три значения усредняются при соблюдении знаков. Результат подобной операции называется иногда сводной взаимокорреляцией рангов; однако его можно было бы более строго назвать по причинам, указанным ниже, коэффициентом соответствия (конкордации). Коэффициент конкордации используется для измерения степени согласованности двух или нескольких рядов проранжированных значений переменных.

В таблице 4.11 представлен результат ранжирования тремя экспертами каких-то условных предметов. Средняя величина из трех значений r_s показывает лишь умеренную степень соответствия. Теперь предположим, что

имеются три значения r_s: 1; –1 и –1. Средняя величина из трех полных корреляций равна вовсе не 1,00, как можно было бы наивно полагать, а всего лишь – 0,33. Все это становится понятным только тогда, если рассматривать эту среднюю как выражение степени соответствия, а не корреляции. В данном случае, при трех полных корреляциях, преобладающим отношением является отношение несоответствия при одном полном соответствии.

Таблица 4.11

Результат ранжирования экспертами неких «условных параметров»

Эксперты	*1 и 2*	r_s =1
Эксперты	*2 и 3*	r_s = –1
Эксперты	*1 и 3*	r_s = –1

Для табличных данных $\overline{r}_s$ = – 0,33.

Хотя усреднение значений показателей любого типа обычно сопряжено с возможными ошибками, в данном случае оно оправдано по той причине, что все значения r_s имеют равный вес, то есть все они вычислены на одном и том же числе наблюдений. Следует повторить, однако, что сама эта средняя величина не является коэффициентом корреляции; она есть средняя из нескольких коэффициентов. Эта величина никогда не может принять значение «минус единица» по той простой причине, что если два ряда коррелируют между собой со значением –1, то третий не может коррелировать с ними обоими одинаковым образом. Однако средняя величина может стать равной «плюс единица» в случае полного соответствия между всеми рядами.

Когда усредняется очень много значений r_s, то вычисления становятся все более трудоемкими. Тем не менее разработан очень простой метод усреднения значения r_s; этот метод особенно полезен, когда число сравниваемых рядов велико. Формула эта не так громоздка, как кажется на первый взгляд, поскольку все символы в ней являются общепринятыми величинами.

$$W = \frac{12S}{k^2 \cdot N \cdot (N^2 - 1)} \quad ,$$

где
- k – число переменных;
- N – число индивидов или категорий, которые ранжируются;
- S = (*Сумма рангов по строке* ***минус*** а)^2, *а* – среднее из суммы рангов.

В приведенной ниже таблице 4.12 демонстрируется применение этой формулы.

Таблица 4.12

Пример вычисления множественного коэффициента ранговой корреляции (коэффициента конкордации W)

Респондент	*Удовлетворенность по признакам А, Б, В*			*Сумма рангов*
	А	***Б***	***В***	
1-й	1	2	1	4
2-й	3	4	5	12
3-й	5	5	4	14
4-й	4	3	3	10
5-й	2	1	2	5
$N=5$				$\Sigma = 45$

Для данных таблицы $a = 45/5 = 9$;

$$S = (4-9)^2 + (12-9)^2 + (14-9)^2 + (10-9)^2 + (5-9)^2 = 76;$$

$$W = \frac{12 * 76}{3^2 * 5 * (5^2 - 1)} = 0{,}84.$$

Значимость полученной величины W для $N>7$ проверяется по критерию X^2: $\chi^2 = \frac{12S}{kN(N+1)}$ со степенью свободы $N-1$. Для рассматриваемого примера $\chi^2_{набл}$ =10,133, степень свободы ($N-1$) = 4. Для α = 0,05 по таблице критических значений χ^2 находим $\chi^2_{кр}$ = 9,488 (см. Приложение 2, *табл. Б*). Поскольку наблюдаемое значение χ^2 больше критического, то необходимо отвергнуть гипотезу H_0 о том, что *не* существует значимой связи между рассматриваемыми переменными.

Полезность r_s. Так как r_s применяется к порядковым данным, не имеющим определенных единиц измерения, этот показатель весьма полезен для социолога, которому часто приходится иметь дело с данными, носящими субъективный характер. Социометрическое ранжирование предпочтений, эстетических суждений и других аналогичных явлений – все это можно коррелировать для того, чтобы определить степень согласия в предпочтениях выборщиков и оценках экспертов. Профессии могут быть ранжированы как по социально-экономическому уровню, так и по уровню разводов. Следовательно, тем самым может быть получена мера предполагаемой взаимосвязи между социально-экономическим уровнем и частотой разводов.

4.8. Линейная корреляция
(для метрических признаков и интервального уровня измерения)

Необходимость общей меры корреляции. Корреляционное поле, которое было тщательно рассмотрено в предыдущем подразделе, позволяет визуально обнаружить совместные вариации двух переменных. Группы изображенных на корреляционном поле точек, которые концентрируются вокруг какой-то гипотетической линии, наводят на мысль об определенном «законе связи» между двумя переменными. Более того, наблюдая ширину разброса, можно сделать, по крайней мере, предварительное заключение о том, насколько хорошо события соответствуют этому гипотетическому закону. Такие заключения зачастую имеют большое значение, но все же оставляют желать лучшего, поскольку субъективны и нестандартизованы. Следовательно, их нельзя как-либо описать или связать с чем бы то ни было, не воспроизводя корреляционного поля. А поскольку эти «визуальные критерии» не точны в математическом смысле, их сопоставление невозможно, даже если в распоряжении исследователя имеются все корреляционные поля во всей их сложности. Поэтому необходим объективный, стандартный, синоптический критерий связи между двумя переменными, конкретная мера корреляции.

То изображение, которое мы видим на корреляционном поле, является гипотетической линией тенденции, которая более или менее отображает совокупность точек, и на основании которой оценивается степень корреляции. Но для того чтобы измерить эту корреляцию, необходимо:

а) точно установить положение такой линии;

б) измерить степень согласия событий, которые ее составляют;

в) посредством специального метода перевести этот результат в индекс корреляции.

Итак, прежде всего, следует позаботиться об определении расположения линии, которая наилучшим образом аппроксимирует наблюдаемые данные.

Напомним, что по типу корреляционная связь может быть прямой или обратной. Прямая связь означает, что при увеличении значения одного признака в среднем увеличивается значение другого, а обратная − при увеличении одного признака в среднем уменьшается значение другого.

По форме корреляционная связь может быть прямолинейной или криволинейной. Гипотезу о форме связи устанавливают по корреляционному полю. ***Прямолинейную*** форму связи имеет такая связь, при которой с увеличением фактора результирующий признак увеличивается или уменьшается на одну и ту же величину. ***Криволинейная*** связь наблюдается тогда, когда подобное изменение происходит неравномерно.

Если связь прямолинейна, то ее можно выразить с помощью уравнения прямой $y = kx + b$, если же связь криволинейна, то она может быть выражена посредством любой кривой, как элементарной, так и нет, например, параболы

($y = ax^2 + bx + c$) или гиперболы ($y = a/x + b$).

По тесноте корреляция может быть *тесной* или *слабой*. Под теснотой связи понимается мера, которая показывает, насколько чувствителен результирующий признак к изменениям факторного признака. О тесноте связи можно судить по корреляционному полю. Если точки распределения плотно сконцентрированы вдоль какой-либо кривой, то говорят, что связь ***тесная***, в противном случае делается вывод о ***слабой*** связи между переменными.

Корреляция может быть также *парной* или *множественной*. ***Парная*** связь устанавливается между двумя признаками (факторным и результирующим). ***Множественная*** связь устанавливается между большим количеством факторных признаков и результирующим.

Все характеристики корреляционного анализа, применяемого для объектов, измеренных по интервальной или порядковой шкале для количественных признаков, определяются следующими коэффициентами:

1) коэффициент наклона линии регрессии $R_{y/x}$;

2) коэффициент детерминации – r^2 и, соответственно, коэффициент недетерминированности ($1-r^2$);

3) коэффициент корреляции Пирсона – r;

4) корреляционное отношение η^2.

Уже говорилось ранее, что связь может быть прямолинейной или криволинейной, следовательно, линия наилучшего приближения может быть прямой или кривой. В данном подразделе ограничимся рассмотрением только линейных связей, то есть таких, которые могут быть представлены в виде прямой линии.

После того как построено корреляционное поле, можно начертить прямую линию, которая при ближайшем рассмотрении оказалась бы линией основной тенденции. Если бы была проведена такая линия, то ее расположение можно было бы обосновать тем, что проведена она, как раз, через середину самой «густой» части скопления всех точек (значений), то есть по возможности, наиболее близко ко всем точкам в среднем, разрезая совокупность этих точек на две одинаковые по длине и ширине полосы. Линия была прочерчена вдоль середины, потому что интуитивно чувствовалось, что именно таким образом ее можно более всего приблизить ко всем наблюдаемым точкам в среднем. Таким образом, она служит оценкой для значений y_i переменной ***Y*** и учитывает каждое наблюдаемое значение x_i переменной ***X***. Оценочные значения y_i соответствуют наблюдаемым значениям x_i и составляют вышеупомянутую линию тенденции. Естественно, чем ближе расположение наблюдаемых точек к оценочной линии, то есть – чем меньше они отклоняются от нее, тем меньше ошибка предсказания y_i по отношению к x_i. Если, например, доход был бы

единственным детерминантом социального статуса, то наблюдаемые точки, все без исключения, обязательно расположились бы на линии соответствующего корреляционного поля; если бы только лишь социальный статус определял доход, то все точки снова попали бы на данную линию.

Итак, какое же значение для корреляции имеют эти отклонения от гипотетической линии регрессии? Они указывают социологу на то, что помимо дохода существуют другие факторы, определяющие социальный статус. Действие этих других, неизвестных, факторов искажает предсказание о природе явления. Как следствие, корреляционная связь между двумя этими переменными оказывается слабой.

Отклонения от среднего арифметического. Сравнивая значения переменных при корреляции, исследователь должен привыкнуть мыслить в терминах отклонений от средних арифметических, вместо того, чтобы анализировать необработанные данные. Подобно многим другим статистическим операциям, которые могли бы, на первый взгляд, показаться излишне косвенными и сложными, эта мера также согласуется со здравым смыслом. Обычно не сравнивают сырые, необработанные результаты наблюдений (рождаемость, жалованье и т. д.), а предпочитают оценивать их как «высокие» или «низкие», то есть как такие, которые лежат выше или ниже среднего показателя или привычной нормы. Однако нельзя сравнивать единицы различных категорий, таких как доход и рождаемость, в их исходной форме, поэтому связывают доход ниже среднего с размером семьи, который превышает среднее.

Таким образом, видно, что среднее арифметическое является удобной точкой начала отсчета, естественным стандартом, в случае, когда сравнивают два ряда данных в отношении их близости друг к другу.

Объясняемые и необъясняемые остаточные отклонения. Проведенная от руки строго через середину линия тенденции, очевидно, является средней для всех наблюдаемых точек, непрерывной последовательностью средних значений, как бы приспособленной к всевозможным значениям случайных событий. Поскольку она состоит из ожидаемых отклонений, значения на наклонной линии тенденции в различных случаях называют «ожидаемыми», «предсказуемыми» или «оценочными» величинами. Так, для примера, рассматривающего взаимосвязь количества высших учебных заведений III–IV уровня аккредитации Украины и численности студентов в них, можно установить на языке статистики, что те отклонения, которые могут быть больше или меньше, чем наблюдаемые, объясняются количеством вузов или могут быть приписаны действию фактора количества вузов. Поэтому данные отклонения получили название объясняемых отклонений.

Поскольку имеются другие факторы, которые влияют на численность студентов (такие, как уровень рождаемости, оплата за обучение и т. д.), то наблюдаемые (действительные, эмпирические) величины не совпадают с ожидаемыми (теоретическими). Воздействие этих неизвестных факторов

измеряется расхождениями между наблюдаемыми и ожидаемыми отклонениями. Чем меньше эти расхождения, тем слабее должно быть воздействие посторонних факторов, чем больше эти расхождения, тем сильнее должно быть это воздействие. Расхождения такого рода иногда называют остатками, которые необъяснимы на основании используемых данных, то есть они должны быть приписаны не количеству вузов, а неизвестным факторам, о сущности которых можно лишь делать предположения. Поэтому их называют необъяснимыми остаточными отклонениями; они также измеряются относительными расстояниями от линии тенденции до отдельных точек исследуемой совокупности. Чем ближе соответствие между объясняемыми и наблюдаемыми отклонениями, то есть чем меньше остатки, тем выше степень корреляции *X* и *Y*. Следовательно, было бы логично измерять степень корреляции в зависимости от степени соответствия между наблюдаемыми и ожидаемыми отклонениями или в виде той доли полного наблюдаемого отклонения, которая объяснима. В этом и заключается основной принцип измерения корреляции, который всегда подсознательно используется при ее вычислении. Конечно, измерение можно было бы произвести с помощью линии регрессии, проведенной от руки, или путем графического анализа при помощи циркуля и линейки, однако существенным остается сравнение, или сопоставление, *полного* и *объясняемого* отклонений, а также нахождение необъясняемого остатка.

Поскольку эти понятия играют такую большую роль в последующем изложении, резюмируем их смысл более точно: 1) ***полное*** (общее или суммарное) наблюдаемое отклонение – это отклонение наблюдаемой величины от среднего арифметического значения; 2) ***объясняемое*** отклонение – это отклонение ожидаемого (регрессионного) значения от среднего арифметического значения; 3) ***необъясняемый остаток*** – это расхождение между полным и объясняемым отклонениями. Разумеется, это есть разность между вышеупомянутыми полным и объясняемым отклонениями.

Измерение линейной корреляции. Серьезный недостаток графиков, выполненных от руки, состоит в том, что они зависят от индивидуального суждения. Без использования стандартной техники вычислений маловероятно, чтобы два исследователя когда-нибудь расположили линию тенденции одинаковым образом. Очевидно, линия, которая используется для измерения корреляции предложенным выше способом, должна всегда удовлетворять одним и тем же требованиям, в противном случае результатам будет недоставать надежности, которая существенна в научной работе.

Одно из таких требований было сформулировано следующим образом: линия располагается так, чтобы сделать равной нулю сумму вертикальных отклонений от этой линии, то есть сбалансировать суммы положительных и отрицательных расхождений. Таким образом, эта линия изображает рассеяние вокруг нее точно так же, как среднее арифметическое представляет совокупность в целом. И подобно среднему арифметическому, она

минимизирует сумму квадратов отклонений, что соответствует принципу наименьших квадратов. Линию следует расположить так, чтобы минимизировать сумму квадратов горизонтальных отклонений, или $X-$ остатков. Однако так как эта линия дала бы ту же меру корреляции, что и линия, минимизирующая квадраты вертикальных отклонений, то необязательно чертить их обе. Для измерения линейной корреляции между двумя переменными достаточно и одной линии регрессии. Имея в виду эти свойства, ее называют линией наилучшего приближения; согласно критерию наименьших квадратов, она приближает рассеяние лучше, чем любая другая прямая. Проведение этой математической линии наилучшего приближения упрощается изображением событий как отклонений от средних арифметических. При соблюдении этих условий линия наименьших квадратов всегда будет проходить через начало отсчета, которое лежит на пересечении средних. Следовательно, ее легко вычертить, как только будет определен ее наклон, – тангенс угла наклона прямой по отношению к оси X.

Вычисление наклона линии регрессии. Вычисление наклона этой линии соответствует следующему уравнению, представленному здесь без объяснения:

$$R_{y/x} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2},$$

где
- $R_{y/x}$ – наклон линии ожидаемых значений в зависимости от X (читается: «Ry по X»);
- $\bar{x}$ – среднее арифметическое по признаку *X;*
- $\bar{y}$ – среднее арифметическое по признаку *Y*;
- $\sum(x-\bar{x})(y-\bar{y})$ – сумма произведений парных отклонений значений переменных ***X*** и ***Y*** (читается: «сумма парных произведений»).

Для примера вычисления коэффициента, показывающего наклон линии регрессии, воспользуемся таблицей динамики сети высших учебных заведений, к которой мы неоднократно обращались в предыдущих подразделах, в частности, посвященных вычислению среднего квадратического отклонения (см. *Табл. 4.13*).

Таблица 4.13

Динамика сети высших учебных заведений Украины III–IV уровня аккредитации и численности студентов в них

Количество высших учебных заведений Украины III–IV уровня аккредитации (x_i)	*Численность студентов в них (тыс.)* (y_i)	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$	$(x_i - \overline{x})^2$
158	718,8	−10568,4	7766,016
159	680,7	−7129	7590,766
232	645	−651,516	199,5156
255	617,7	−167,0719	78,76563
274	595	−108,016	777,0156
280	526,4	−2455,09	1147,516
298	503,7	−4937,2	2691,016
313	503,7	−6364,83	4472,266
$\sum$ 1969	4791	−32047	24722,88
$\overline{X}$ = 246,125	$\overline{Y}$ = 598,875		

Это отношение задает наклон искомой линии наилучшего приближения и представляет собой среднее изменение значения y_i при изменении значения x_i на единицу. Соответственно, чтобы найти среднее изменение в численности студентов на единичное изменение в количестве вузов, вычисляют парные произведения, возводят в квадрат отклонения по X и образуют отношение между их соответствующими суммами. Все эти действия отображены в таблице 4.13, представленной выше. Осуществив соответствующие вычисления, мы получили $R_{y/x}$ = –1,296. Это и есть средняя скорость изменения численности студентов на единицу изменения количества вузов, то есть при каждом изменении количества вузов на единицу численность студентов уменьшается в среднем на 1,296 тыс. человек. Поскольку эта величина фиксирует положение линии регрессии относительно оси независимой переменной (в нашем случае – оси X), она называется наклоном. Как только наклон определен, можно вычертить линию наименьших квадратов и перейти к точному измерению корреляции.

Коэффициент детерминации.. Существует несколько альтернативных определений коэффициента детерминации, однако в случае линейной регрессии

все они эквивалентны. ***Коэффициент детерминации*** − это показатель того, насколько изменения зависимого признака объясняются изменениями независимого. Если более точно− это доля дисперсии независимого признака, объясняемая влиянием зависимого. Вычисление этого коэффициента основывается на том очевидном принципе, что чем ближе объясняемые отклонения к общей вариации, тем больше доля объясняемой вариации и тем выше степень корреляции. Чем значительнее доля объясненной вариации, тем меньше роль прочих факторов.

Как статистическая операция, преобразование этих отклонений в отдельный индекс состоит в суммировании квадратов объясняемых отклонений и выражает вариацию как долю обшей вариации, которую необходимо объяснить. Символически это можно представить так:

$$r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2},$$

где
- $\hat{y}_i$ – ожидаемое значение зависимой переменной, предсказанное по уравнению регрессии;
- y_i – наблюдаемое (эмпирическое) значение зависимой переменной;
- $\bar{y}$ – среднее значение зависимой переменной.

Коэффициент детерминации изменяется в диапазоне от 0 до 1. Соответствующее неравенство имеет следующий вид: $0 \leq r^2 \leq 1$

При этом, если $r^2 = 0$ − это означает, что связь между переменными регрессионной модели отсутствует, и вместо нее для оценки значения выходной переменной можно с таким же успехом использовать простое среднее ее наблюдаемых значений. Когда же $r^2 = 1$ − имеет место соответствие идеальной модели, когда все точки наблюдений лежат точно на линии регрессии, то есть сумма квадратов их отклонений равна 0. На практике, если коэффициент детерминации близок к 1, это указывает на то, что модель работает очень хорошо (имеет высокую значимость), а если − к 0, то это означает низкую значимость модели, когда независимая переменная (***X***) плохо «объясняет» поведение зависимой переменной (***Y***), то есть линейная зависимость между ними отсутствует. Очевидно, что такая модель будет иметь низкую эффективность.

Другими словами, r^2 представляет собою долю в общей вариации зависимой переменной ***Y***, которая обусловлена изменением независимой переменной ***X***, поэтому r^2 и называют коэффициентом детерминации. Так как объяснить можно не более, чем всю полную вариацию, то r^2 никогда не может превышать единицу, а практически всегда бывает меньше ее.

Полное вычисление r^2 показано в таблицах 4.14 и 4.15, где и объясняемые и наблюдаемые отклонения возводятся в квадрат и суммируются,

что дает соответственно объясняемую и общую вариацию.

Таблица 4.14

Распределение взаимных частот, описывающих связь дохода в сотнях гривень и оценкой социального статуса

Доход в сотнях гривень (X)	*Оценка социального статуса* (Y)	$(X_i - \overline{X})$	$(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
3	3	–7	–3	49	21
7	5	–3	–1	9	3
11	7	1	1	1	1
14	6	4	0	16	0
15	9	5	3	25	15
$\overline{X}$=10	$\overline{Y}$=6	0	0	**100**	**40**

Исходя из приведенной выше формулы для определения наклона линии регрессии, в данном случае $R_{y/x} = \frac{40}{100} = 0{,}4$. Значит, уравнение теоретических распределений выглядит так $\hat{y} - \overline{y} = R_{y/x}(x_i - \overline{x})$.

Таблица 4.15

Продолжение вычисления для таблицы 4.14

$\hat{y} - \overline{y}$	$(\hat{y} - \overline{y})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$
0,4*(–7)=–2,8	7,84	–3	9
–1,2	1,44	–1	1
0,4	0,16	1	1
1,6	2,56	0	0
2,0	4,00	3	9
0	**16**	0	**20**

r^2=16/20=0,8

Таким образом, деля объясняемую вариацию 16 на общую вариацию 20, мы получили r^2= 0,8. Это и есть окончательная мера степени связи между двумя переменными: она показывает, что на 80% вариация переменной ***Y*** объясняется линейной зависимостью от переменной ***X***.

Обратимость r^2. С одинаковым успехом могли бы вычислить как регрессию ***X*** от ***Y***, так и регрессию ***Y*** по ***X*** и тем самым определить долю вариации ***X***, объясняемую через ***Y***. Действительно, такой результат был бы существенен для окончательной формулировки корреляции. Почему же в таком

случае вычисляют r^2_{xy} (***X*** по ***Y***)? Ответ состоит в том, что вклад r^2_{xy} учитывается автоматически. Нет необходимости строить линии регрессии дважды по той причине, что $r^2_{yx} = r^2_{xy}$. Короче говоря, r^2 обратимо. Например, зная, что $r^2_{yx} = 0{,}80$, можно утверждать не только, что доход объясняет 80% вариации в социальном статусе, но также и наоборот, а именно: социальный статус объясняет 80% вариации в доходе. Так как с точки зрения статистики одно объясняет другое в одинаковой степени, то подписные значки при r^2 обычно опускают.

Коэффициент недетерминированности. Поскольку разность между общей вариацией и объясняемой вариацией обязательно равна необъясняемой вариации, то необъясняемая доля вариации есть просто разность от единицы и соответственно называется ***коэффициентом недерминированности***. Его можно записать следующим образом:

$$(1 - r^2) = 1 - \frac{\Sigma_i (\hat{y}_i - \bar{y})^2}{\Sigma_i (y_i - \bar{y})^2},$$

где
- $\hat{y}_i$ – ожидаемое значение зависимой переменной, предсказанное по уравнению регрессии;
- y_i – наблюдаемое (эмпирическое) значение зависимой переменной;
- $\bar{y}$ – среднее значение зависимой переменной.

Можно бы получить эту величину непосредственным измерением остатков, расположенных вокруг линии регрессии, возведением их в квадрат и суммированием, выражая эту сумму как долю полной вариации. В действительности, в этом и состоит операциональный смысл коэффициента недетерминированности $1 - r^2$. Но независимо от того, вычисляется ли этот коэффициент непосредственно из остатков или косвенно через величину r^2, его всегда можно истолковать как долю вариации зависимой переменной, которую нельзя линейно объяснить через независимую переменную. Поэтому он измеряет силу воздействия неизвестных факторов.

Коэффициент корреляции Пирсона (r). Важной характеристикой совместного распределения двух случайных величин является ковариация. В теории вероятности ***ковариа́ция*** − это мера линейной зависимости случайных величин. Считается, что если две случайные величины ***X*** и ***Y*** имеют в отношении друг друга линейные функции регрессии, то эти величины (***X*** и ***Y***) ковариантны, а значит, связаны между собой линейной корреляционной зависимостью. Если двумерная случайная величина (***X, Y***) распределена нормально, то между ***X*** и ***Y*** имеет место линейная корреляция.

Если величины ***X*** и ***Y*** независимы, то **cov (*X,Y*)** = 0. Такие величины, то есть те, которые имеют нулевую ковариацию, называются

некоррелированными. Ковариация – мера связи, широко применяемая в физике и технических науках, которая определяется по следующей формуле:

$$COV_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N - 1},$$

где
- x_i – значения, принимаемые переменной ***X***;
- y_i – значения, принимаемые переменой ***Y***;
- $\bar{x}$ – среднее значение по ***X***;
- $\bar{y}$ – среднее значение по ***Y;***
- ***N*** – число единиц совокупности.

Однако в социологии, в отличие от физики, например, большинство переменных измеряется в произвольных шкалах. Ковариация служит характеристикой взаимозависимости случайных величин, однако только лишь по ее абсолютному значению нельзя судить о том, насколько сильно величины взаимосвязаны, так как величина ковариации зависит от единиц измерения независимых величин. Данная особенность ковариации затрудняет ее использование в целях корреляционного анализа. Для устранения недостатка ковариации, для того, чтобы сделать меру связи независимой от единиц измерения того или иного признака, достаточно разделить ковариацию на соответствующие стандартные отклонения [4, с. 70]. Таким образом и была получена формула ***коэффициента корреляции Пирсона***, который разработали Карл Пирсон, Фрэнсис Эджуорт и Рафаэль Уэлдон в 90-х годах XIX века. Этот коэффициент рассчитывается по следующей формуле:

$$\boldsymbol{r} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(N-1)\sigma_x \sigma_y}$$

или

$$\boldsymbol{r} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \times \sum_i (y_i - \bar{y})^2}},$$

где
- x_i – значения, принимаемые переменной ***X***;
- y_i – значения, принимаемые переменой ***Y***;
- $\bar{x}$ – среднее значение по ***X***;
- $\bar{y}$ – среднее значение по ***Y***;
- σ_X и σ_Y – выборочные средние квадратические отклонения величин ***X*** и ***Y***.

Последняя формула является основной для вычисления коэффициента корреляции Пирсона. Этот коэффициент показывает тесноту линейной связи между ***X*** и ***Y*** : чем ближе | ***r*** | к единице, тем сильнее линейная связь между ***X*** и

Y. Следовательно, r изменяется в пределах от 1 до 1, что может быть представлено следующим неравенством: $-1 \le r \le 1$.

В качестве примера вычисления вновь обратимся к данным таблицы 4.14 уже рассматриваемой ранее в этом подразделе.

Таблица 4.16

Вычисление r для данных таблицы 4.14 «Распределение взаимных частот, описывающих связь дохода в сотнях гривень и оценкой социального статуса»

Доход в сотнях гривень (X)	*Оценка социально-го статуса (Y)*	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
3	3	3−10=−7	49	3−6=−3	9	(−7)×(−3)=21
7	5	−3	9	−1	1	3
11	7	1	1	1	1	1
14	6	4	16	0	0	0
15	9	5	25	3	9	15
$\overline{X}$=10	$\overline{Y}$=6		$\sum_i = 100$		$\sum_i = 20$	$\sum_i = 40$

Подставив зафиксированные в таблице 4.16 цифры в основную расчетную формулу для нахождения коэффициента корреляции Пирсона, получим:

$$r = \frac{40}{\sqrt{100 \times 20}} = \frac{40}{44{,}72} = 0{,}89 .$$

Такое значение коэффициента свидетельствует о наличии весьма сильной прямой связи между уровнем дохода респондентов и их оценкой собственного социального статуса.

В буквальном смысле величину r можно истолковывать как среднее изменение Y на каждое единичное изменение X, или наоборот, предполагая в обоих случаях, что критерии выражены в стандартной форме, то есть в долях σ. Итак, если известно, что доход отклоняется на величину 1σ от среднего, то следует ожидать, что соответствующий ему уровень социального статуса отклоняется в среднем от своего среднего арифметического значения на $0{,}89\sigma$.

На основании данного принципа можно оценить Y (в стандартной форме) для любого данного значения X точно так же, как это было сделано для получения ожидаемых отклонений Y. Следует лишь применить формулу: $\hat{z}_y = r(z_x)$, где $\hat{z}_y$ – оценочное значение в единицах σ для величины X. Подставляя в формулу данные, получаем оценочные величины, приведенные в таблице 4.17.

Таблица 4.17

Оценка Z_y по наблюдаемому Z_x

Доход в сотнях гривень (X)	*Оценка социального статуса (Y)*	Z_x	$\hat{z}_y = r(z_x)$
3	3	–1,56	–1,39
7	5	–0,67	–0,60
11	7	0,22	0,20
14	6	0,89	0,79
15	9	1,12	1,00
$\overline{X}$=10	$\overline{Y}$=6		

Итак, когда доход в стандартной форме составляет $1{,}12\sigma$, то социальный статус оценивается как $1{,}00\sigma$, если доход равен $-\ 1{,}56\sigma$, то социальный статус $-\ 1{,}39\sigma$. Полная корреляция между двумя рядами данных имела бы место всякий раз, когда парными значениями были бы идентичные расстояния в долях σ от соответствующих средних. Например, если бы человек с оценкой по социологии, скажем, на $1{,}8\sigma$ выше среднего, располагал бы на $1{,}8\sigma$ выше среднего значения для оценок по истории, то и все другие наблюдения в этих двух рядах были бы совершенно аналогичными. Поскольку r обратим, можно применить его к наблюдаемым значениям Z_y и тем самым оценить соответствующую величину Z_x.

В результате осуществленных вычислений, описанных выше, очевидность приобретает тот факт, что основная формула вычисления коэффициента корреляции Пирсона, хотя и вполне понятна, однако, мягко говоря, не очень удобна для вычисления «вручную», даже с использованием калькулятора. Поэтому существуют производные формулы – более громоздкие по виду, менее доступные осмыслению и не совсем понятные, однако существенно упрощающие расчеты. Кроме того, эти формулы, в свою очередь, дифференцируются на: «для сгруппированных данных» и «для несгруппированных данных». В данном издании мы не будем их приводить, так как «вручную» коэффициент корреляции вычисляется крайне редко (только в каких-то крайних случаях), а в основном для обработки реальных данных социологического исследования с целью определения корреляционных связей и зависимостей используются специальные компьютерные программы.

Основные свойства коэффициента корреляции Пирсона. Как мера тесноты и направления (линейной) связи между двумя признаками, коэффициент корреляции Пирсона r обладает следующими свойствами:

1) $r_{y/x} = r_{x/y}$ (поэтому, как правило, этот коэффициент обозначается просто одной буквой r);

2) чем ближе $r_{y/x}$ к +1 или –1, тем теснее связь. Чем ближе $r_{y/x}$ к 0, тем связь слабее. Если $r_{y/x} = 0$, то говорят, что связи нет. Для социальных процессов $r_{y/x}$ редко превышает |0,75|;

3) если $r_{y/x} > 0$, то связь прямая. Если $r_{y/x} < 0$, то связь обратная.

Сравнение r и r^2. При поверхностном рассмотрении разница между коэффициентами r и r^2 кажется весьма тривиальной: налицо просто разные показатели степени, причем обе величины легко преобразуются друг в друга, как только одна из них вычислена.

Если рассмотреть такие преобразования на примере связи дохода респондентов и их оценки собственного социального статуса (см. *Табл. 4.16* и *4.17*) при коэффициенте корреляции Пирсона $r = 0{,}89$ получим коэффициент детерминации $r^2 = 0{,}79 \approx 0{,}8$ (что, кстати, вполне согласуется с результатами вычислений по нахождению r^2, осуществленных с применением специальной формулы).

Тем не менее этой кажущейся незначительной разницей нельзя с легкостью пренебрегать, потому что оба коэффициента используются в двух различных, хотя и взаимосвязанных, аспектах ковариации.

Каждая из двух рассматриваемых величин механически выводится из другой посредством одной из двух процедур. Однако критерий r^2, полученный из r, не дает полного представления о назначении и смысле объясняемой вариации, которая измеряется величиной r^2. Понятие «квадрат наклона» ничего не говорит студенту. В такой формулировке, как «корень квадратный из объясняемой вариации», нельзя было бы усмотреть наклон. Хотя r и r^2 взаимозависимы, критерий r^2 измеряет всю ту долю полной вариации одной переменной, которая связана с другой или объясняется ею.

С другой стороны, критерий r оценивает динамический аспект этого отношения, измеряя скорость изменения одной переменной относительно другой, как это было показано в предыдущих примерах. Исходя из этого концептуального различия, можно утверждать, что критерий r является, главным образом, средством предсказания, например, ожидаемого уровня изменения одной переменной при наблюдаемом изменении другой. Как таковой он пригодился бы педагогам, психологам и другим исследователям, которых интересует единичное предсказание, а вот социологам – в значительно меньшей степени. С другой стороны, r^2 есть суммарная мера, взвешивающая влияние (или воздействие) одной переменной на другую.

Поскольку величина r представляет собой наклон, то она, очевидно, должна задавать направление преимущественно вверх или вниз в соответствии

с тем, положительно или отрицательно связаны переменные. Это означает, что направление наклона отражает тип связи, который обозначается знаком «плюс» или «минус». Синоптическая мера r^2, однако, не несет знака, ибо она выражает долю общей вариации.

Теперь ясно, что r и r^2 не являются взаимозаменяемыми; их также не следует непосредственно выводить друг из друга до тех пор, пока их структурный смысл не будет понят. Так как значение r всегда больше, чем значение r^2, то случайное или преднамеренное использование r для выражения силы связи вводило бы читателя в заблуждение: например, могло бы показаться, что $r = 0{,}5$ означает довольно сильную связь, но $r^2 = 0{,}25$ показывает, что лишь 25% вариации каждой переменной связано с другой. Когда требуется подчеркнуть силу полной связи между двумя переменными, что часто бывает в социологических исследованиях, более подходящей статистикой является r^2. Например, если выяснено, что 50% вариации в области преступности может быть связано с экономическим фактором, хотя кое-что остается еще необъясненным (какими факторами вызваны еще 50% вариации в области преступности), все же это является реальным продвижением на пути понимания явления, которое всего сто лет назад объясняли нечистой силой, наследственностью или свободной волей.

В итоге каждый критерий имеет свое собственное обозначение и соответствующее употребление. Тем не менее, критерий r пользуется большей популярностью по сравнению с r^2. Возможно, отчасти это объясняется силой привычки, которая, вероятно, исчезнет, когда исследователи станут более чувствительными к нюансам количественного описания.

Делая общий вывод, подчеркнем, что используя коэффициент корреляции Пирсона, следует учитывать, что лучше всего он подходит для оценки взаимосвязи между двумя переменными, значения которых распределены нормально (т. е. согласно закону нормального распределения). Если распределение переменных отличается от нормального, то этот коэффициент по-прежнему продолжает характеризовать степень взаимосвязи между признаками, однако в данном случае к нему уже нельзя применять методы проверки на значимость. Также коэффициент корреляции Пирсона не очень устойчив к выбросам (резко выдающимся значениям). То есть в тех случаях, когда есть резко выделяющиеся значения, можно ошибочно сделать вывод о наличии корреляции между переменными. Поэтому если распределение исследуемых переменных отличается от нормального или возможны выбросы, то лучше воспользоваться либо «непараметрическим аналогом», то есть коэффициентом ранговой корреляции Спирмена (о котором подробно говорилось в предыдущем подразделе), либо обратиться к другим методам определения корреляции, один из которых будет подробно описан далее.

4.9. Нелинейная регрессия. Множественная и частная корреляция

Нелинейная регрессия. Всякий коэффициент корреляции показывает нам, в какой степени одну переменную можно предсказать или объяснить через другую. Из многих употребляемых в статистике показателей коэффициент r Пирсона является одной из наиболее распространенных, почти банальных мер. Однако применимость r основывается в основном на следующих двух условиях: 1) отношение между переменными линейно; 2) двумерное распределение гомоскедастично, то есть имеет постоянную условную дисперсию. Вообще говоря, оба эти условия выполняются, если маргинальные распределения нормальны. Поэтому одним из требований, предъявляемых к критерию r, часто является нормальность маргинальных распределений. Однако очень часто исходные данные приводят к нелинейным моделям. Это особенно справедливо для социологических наук, где многие распределения, такие как доход или размер семьи, сильно скошены, и линии регрессии, поэтому, в значительной мере нелинейны.

Так как критерий r определяет прямолинейную модель, то становится неподходящим в той мере, в какой связь между переменными отклоняется от линейной. То же самое получается, если диаграмма рассеяния гетероскедастична, даже если при этом она остается линейной. Конечно, поскольку эти идеальные условия никогда не могут быть выполнены, исследователи вынуждены прибегать к аппроксимации. Но существует разумный предел, за который не следовало бы распространять аппроксимацию, если доступны более подходящие способы измерения.

Существуют различные методы, с помощью которых можно измерить криволинейные зависимости, но здесь будем рассматривать только один из них. Излагаемый метод есть метод корреляционного отношения, который часто обозначается через η (греческая строчная буква «эта»), но лучше выражается посредством η^2. Хотя обычно «корреляционное отношение» отождествляют просто с η, во время обсуждения соответствующей темы, однако, станет ясно, что критерий η мало пригоден для практического применения; именно поэтому он является незначимым критерием корреляции. Поэтому термин «корреляционное отношение» будет использоваться как равнозначный с величиной η^2.

Данный метод представляет собой относительно простую процедуру и связан с тем же принципом, на котором основано определение r^2, а именно: на соотношении объясняемой и общей вариации. Вычисления в этом случае отличаются от определения r^2 в том отношении, что объясняемая вариация выводится из среднего значения по столбцам (строчкам) корреляционной таблицы, а не из гипотетической линии регрессии. Тем самым корреляционное

отношение дополняет «пирсоновский» критерий r^2, пользующийся популярностью и репутацией, вероятно, превышающими его практическую значимость в области социологических наук.

Вычисление корреляционного отношения. В целях простой иллюстрации вычислим корреляционное отношение между численностью населения Украины и годами с 1986 по 1999 по официальным данным статистики (см. *Табл. 4.18*)

Таблица 4.18

Численность населения Украины в 1986–1999 годах

Год	*Численность (млн человек)*
1986	51
1991	51,9
1992	52,1
1993	52,2
1994	52,1
1995	51,7
1996	51,3
1997	50,9
1998	50,5
1999	50,1

Чтобы установить, соответствуют ли рассматриваемые данные линейной или нелинейной модели, надо, прежде всего, построить обычное корреляционное поле (см. *Рис. 4.7*), которое позволит определить это визуально.

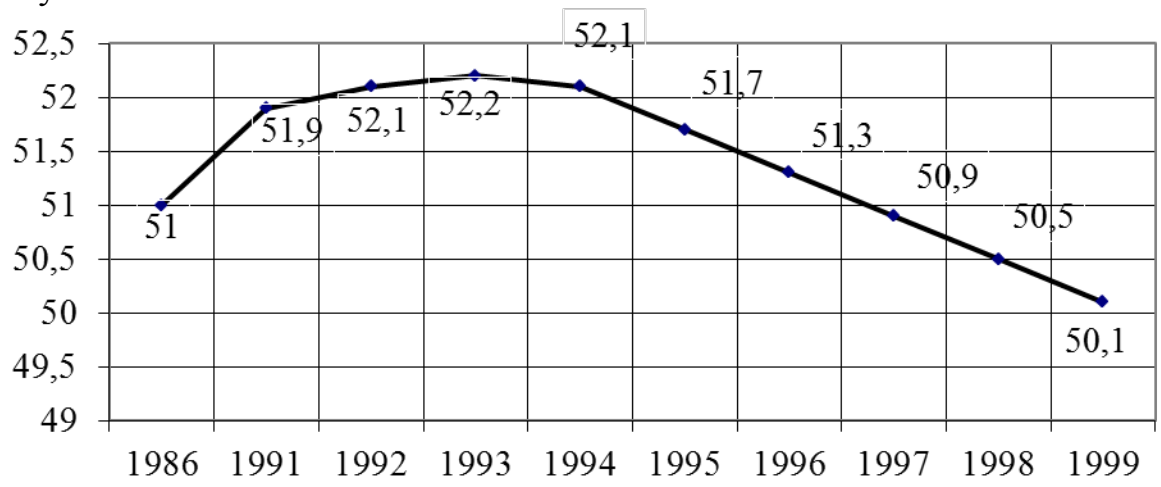

Рис. 4.7. Численность населения Украины в 1986–1999 годах: корреляционное поле

Корреляционное поле, на которое наложена проведенная от руки линия тенденции, позволяет оценить силу и тип связи между двумя переменными. Можно было бы представить разброс и прямой линией, но кривая следует контуру данных более точно. Поэтому заключаем, что η^2 дает лучшее приближение, чем r^2, и приступаем к вычислению η_{yx} .

Процедура вычисления значения η^2 совершенно аналогична той, что имела место при вычислении r^2 Пирсона. Необходимо вычислить полную вариацию и объясняемую вариацию, а затем найти отношение между ними. Соответствующая формула (для несгруппированных данных) выглядит следующим образом:

$$\eta^2_{yx} = \frac{объясняемая\_вариация}{полная\_вариация} = \frac{\sum_{j=1}^{k}(\hat{y}_i - \overline{y})^2}{\sum_{j=1}^{N}(y_i - \overline{y})^2},$$

где
- k − число групп по факторному признаку;
- *N* − число единиц совокупности;
- $\hat{y}_i$ − внутригрупповое среднее значение признака ***Y***;
- $\overline{y}$ − общее среднее значение признака ***Y***;
- y_i - индивидуальные значения признака ***Y***.

Данная формула представляет собой вычисление η^2.

В случае, если данные сгруппированы, эта формула преобразовывается и имеет такой вид:

$$\eta^2_{yx} = \frac{объясняемая\_вариация}{полная\_вариация} = \frac{\sum_{j=1}^{k}(\hat{y}_i - \overline{y})^2 \cdot n_j}{\sum_{j=1}^{N}(y_i - \overline{y})^2},$$

где
- k − число групп по факторному признаку ***X***;
- *N* − число единиц совокупности;
- $\hat{y}_i$ − внутригрупповое среднее значение признака ***Y***;
- $\overline{y}$ − общее среднее значение признака ***Y***;
- y_i − индивидуальные значения признака ***Y***;
- n_j − частота значения признака в *j*-ой группе.

Процедуру в целом лучше всего описать по этапам, что мы и сделаем, начиная с таблицы 4.19.

Таблица 4.19

Численность населения Украины в 1986–1999 годах

Год (x_i)		*Численность (млн. человек)* (y_i)	$(y_i-\bar{y})^2$	$\hat{y}$ – *среднее арифметическое – групповое*	$(\hat{y}_i-\bar{y})^2$
1	1986	51	$(51–51,38)^2$=0,1444	(51+51,9)/2= 51,45	$(51,45–51,38)^2$= 0,005
2	1991	51,9	0,2704	51,45	0,005
3	1992	52,1	0,5184	52,133	0,567
4	1993	52,2	0,6724	52,133	0,567
5	1994	52,1	0,5184	52,133	0,567
6	1995	51,7	0,1024	51,5	0,014
7	1996	51,3	0,0064	51,5	0,014
8	1997	50,9	0,2304	50,5	0,77
9	1998	50,5	0,7744	50,5	0,77
10	1999	50,1	1,6384	50,5	0,77
N=10	∑	513,8	**4,876**		**4,049**

1. Применяем обычный способ двумерной группировки данных, в нашем случае – это данные распределения численности населения Украины (признак ***Y***) в 1986–1999 годах (признак ***X***). Результаты группировки, приведенны в таблице 4.19, представленной выше.

2. Вычисляем общее среднее для значений признака ***Y*** по формуле: $\bar{y}=\frac{\Sigma y_i}{N}=\frac{513,8}{10}=51,38\cdot$

3. Подсчитываем сумму квадратов отклонений по y, или общую вариацию по формуле: $\sigma^2=\Sigma(y_i-\bar{y})^2$ =4,876.

4. Вычисляем среднее $\hat{y}$ для значений y в каждой группе, причем разделение на группы определяется исследователем с точки зрения здравого смысла. В нашем случае деление на группы определено – подчеркиванием.

5. Находим сумму квадратичных отклонений. Это и есть объясняемая вариация.

6. Делим объясняемую вариацию на полную вариацию и получаем, согласно определению, корреляционное отношение:

$$\eta^2_{yx}=\frac{объясняемая\_вариация}{полная\_вариация}=\frac{4,049}{4,876}=0,83.$$

Итак, было найдено, что 0,83% вариации численности населения Украины объясняется изменением года. Действуя точно в той же

последовательности, но уже со средними для строк, а не для столбцов, можно определить η^2_{xy}– часть вариации X, объясняемую через Y. Тот факт, что эти два значения η^2 различны, указывает, что, в противоположность r^2, критерий η^2 не является обратимым. В общем случае относительная вариация в строках и столбцах не совпадает в точности, поэтому значения η^2_{xy} чаще всего не равно значению η^2_{yx}. В крайнем, весьма маловероятном случае, вариация по столбцам может отсутствовать и вовсе, оставаясь существенной по строкам, так что значение η^2_{yx} было бы равно единице, а значение η^2_{xy} обнаруживало бы лишь умеренную степень связи.

Сравнение статистических показателей r^2 и η^2:

- $r^2 = 0$, если X и Y независимы.
- $r^2 = \eta^2_{x/y} = 1$, тогда и только тогда, когда имеется строгая линейная связь между X и Y.
- $r^2 \le \eta^2_{x/y}$, тогда и только тогда, когда имеется строгая нелинейная функциональная зависимость X и Y.
- $r^2 = \eta^2_{x/y} < 1$, тогда и только тогда, когда регрессия X и Yстрого линейная, но нет функциональной зависимости.
- $r^2 < \eta^2_{x/y} < 1$ означает, что нет функциональной зависимости и существует нелинейная кривая регрессии.

Условия применимости критерия η^2. Границы применения критерия η^2 более широки, нежели для критерия r^2 Пирсона. Так, не накладывается никаких ограничений на форму маргинальных распределений, и они могут быть нормальными, скошенными или даже бимодальными; линия регрессии может иметь любую криволинейную форму. Как переменную по одной оси можно использовать даже качественную переменную (по двум осям – нельзя, ибо тогда невозможно было бы найти среднее арифметическое и пришлось бы вычислять коэффициент взаимной сопряженности C). Можно предложить только один ограничивающий фактор: совокупность нанесенных на диаграмму точек должна в общем случае быть гомоскедастична, по крайней мере, в одном направлении; это необходимо по той же причине, что и для r Пирсона, а именно потому, что синоптическая мера обычно более значима, если ее получают из однородных данных. Но это – скорее логическое, а не математическое ограничение. По этой же причине иногда нецелесообразно

применять среднее для данных с большой дисперсией.

Принципы интерпретации. Так как гипотетическая линия регрессии меняет направление из-за своей кривизны и изгибов, то весьма трудно интерпретировать величину η^2 тем же способом, как и r^2: «чем выше значние X, тем выше значение Y» (методически: «среднее изменение Y, приходящееся на единицу изменения X»). Такая интерпретация была бы лишена смысла, поскольку линия регрессии может менять направление, следовательно, величина не эквивалентна постоянному наклону r. Действительно, η не выполняет никакой определенной функции в измерении корреляции; она лишь является корнем квадратным из величины η^2, которая больше подходит для нелинейной корреляции.

Поскольку всякая совокупность нанесенных на диаграмму точек всегда будет содержать в себе определенную нелинейность, то значение η^2, отражающее криволинейность и линейность одинаково хорошо, всегда будет более точной оценкой для объясняемой вариации и, следовательно, будет всегда выше, чем значение r^2. Критерий r^2 всегда предельно расширяет возможности статистического объяснения, ибо он автоматически приспосабливается к конфигурации данных. С другой стороны, критерий r^2 является более жестким, что ограничивает его возможности отражением лишь линейной модели. В случае полной линейной зависимости $r^2 = \eta^2$.

Причина, по которой критерий r^2 может более полно отражать объясняемую вариацию, состоит просто-напросто в том, что он делит искривленную линию регрессии на сегменты, которые всегда находятся в большей близости к точкам совокупности, чем одна прямая линия. Поэтому остатки будут меньше. В строго линейной совокупности с одним только направлением такая гибкость, разумеется, не нужна.

На рисунке 4.7 видно, что для данной совокупности подходит не прямая, а именно кривая линия регрессии. Однако в математических методах нет ничего, что исключало бы применение линейной формулы к нелинейным данным. При этом можно было бы уловить в какой-то мере линейность, которая в действительности может вообще отсутствовать. Кривая линия просто лучше приближает данные и поэтому больше подходит для целей исследования.

Большая точность криволинейного приближения становится все очевиднее, когда используется линия регрессии по ее назначению, а именно для предсказания одной переменной по другой. Очевидно, значения Y, определяемые по прямой линии, приведут к значительно большим ошибкам, чем в случае использования кривой, ибо прямая линия становится все менее и менее представительной по мере того, как данные становятся все более и более нелинейными. В самом деле, на краях распределения данные далеко

отклоняются от прямой линии, так что предсказуемые значения сильно отклоняются от наблюдаемых.

Более того, если к нелинейной зависимости применяется линейная модель, нормирующая способность величины r^2 уменьшается, потому что индекс не может меняться от нуля до единицы. Значения r^2, если его ошибочно применять к нелинейным данным, не достигало бы единицы. Но критерий η^2, как это уже было показано, не имеет такого ограничения и теоретически может достигать единицы. В этом случае предполагается, что при определенных условиях, когда r^2 не подходит, то обращаются к η^2.

Предосторожности в применении критерия η^2. Чтобы использовать формулу для η^2, необходимо сгруппировать данные по сегментам, а такая группировка уже влечет за собой в определенной степени произвол. Доводя этот процесс до абсурда, можно выделить столько групп классификации, или интервалов группировки, сколько имеется точек, располагая каждую наблюдаемую точку в отдельном столбце. Кривая регрессии проходила бы тогда через другую точку. При такой группировке значения η^2 было бы в точности равно единице, ибо всякое отклонение внутри строк исключалось бы. С другой стороны, интервалы группировки можно было бы сделать абсурдно большими, и тогда не удалось бы выявить форму распределения. В таком случае остатки были бы слишком велики для того, чтобы судить о кривизне распределения, и мы вновь столкнулись бы с известным явлением неправильной группировки.

Виды нелинейной формы связи. В случае, когда $r^2 \le \eta^2_{x/y}$, между признаками существует строгая нелинейная связь. Задача социолога отыскать уравнение функциональной зависимости, определяющее эту связь, то есть уравнение регрессии. В данном издании представим без особых объяснений только два уравнения криволинейной зависимости: параболическое и гиперболическое.

$\overline{y}_x = a + b \cdot x + c \cdot x^2$ – уравнение, определяющее параболу.

Для нахождение параметров *a, b, c* необходимо решить следующую систему уравнений:

$$\begin{cases} \sum y = na + b\sum x + c\sum x^2 \\ \sum yx = a\sum x + b\sum x^2 + c\sum x^3 \\ \sum yx^2 = a\sum x^2 + b\sum x^3 + c\sum x^4. \end{cases}$$

Для сгруппированных данных формулы поиска параметров *a, b, c* выглядят так:

$$\begin{cases} \sum yn_y = na + c\sum (x-\bar{x})^2 \cdot n_x \\ \sum y(x-\bar{x}) \cdot n_y \cdot n_x = b\sum (x-\bar{x})^2 n_x \\ \sum y(x-\bar{x})^2 \cdot n_y \cdot n_x = a\sum (x-\bar{x})^2 \cdot n_x + c\sum (x-\bar{x})^4 \cdot n_x. \end{cases}$$

В случае если корреляционное поле представляет собой кривую, которая может быть описана гиперболой, применяется формула: $y_x = a + \frac{b}{x}$.

Для нахождения параметров *a, b* необходимо решить систему уравнений:

$$\begin{cases} \sum y/x = a\sum 1/x + b\sum 1/x^2 \\ \sum y = an + b\sum 1/x \end{cases}$$ − для несгруппированных данных;

$$\begin{cases} \sum y/x\, n_n \cdot n_y = a\sum n_x/x + b\sum n_x/x^2 \\ \sum y \cdot n_y = an + b\sum n_x/x \end{cases}$$ − для сгруппированных данных.

Частная и множественная регрессия и корреляция. Вывод о связи может быть сделан только на основании анализа всей совокупности связей в системе «изучаемый процесс – факторные признаки», поэтому рассчитывают не отдельные коэффициенты, а таблицу коэффициентов.

Так, если S_0 − показатель, отражающий уровень интереса студентов к социологии, а F_1, F_2, F_3 − факторы учебного процесса, отражающие содержание программы, уровень квалификации преподавателей и объем программы в часах, матрица коэффициентов связи может выглядеть так:

Таблица 4.20

Матрица коэффициентов связи (эмпирический пример)

	S_0	F_1	F_2	F_3
S_0	1	0,89	0, 68	0,75
F_1	0,89	1	0,56	0,95
F_2	0,68	0,56	1	0,11
F_3	0,75	0,95	0,11	1

Чисто внешне значимы все три фактора, однако при анализе внутренних связей можно заметить, что оценка содержания программы оказалась зависимой от ее объема в часах, поэтому, несмотря на то что уровень связи изучаемого процесса с показателем «содержание программы» выше,

детерминирующим следует считать фактор «объем программы в часах», так как именно он определяет меру содержательности программы. Кроме того, на содержание программы оказывает значимое влияние и фактор квалификации преподавателей.

Построение математической факторной модели включает оценку количественного влияния факторов на изучаемый процесс. Модель разрабатывается после качественного анализа влияния факторов и требует включения только тех факторов, влияние которых на изучаемый процесс доказано на предыдущем этапе. Модель, как правило, представляется в виде регрессионной функции вида: $y = f(x_1, x_2, ...x_k)$. Вид функции выбирается исходя из качественного анализа процесса или подбирается путем перебора. Для моделирования вида функции часто используют стандартные пакеты типа Statgraf или Statistica для Windows.

Ранее было показано, как можно по опытным данным найти зависимость одной переменной от другой, а именно, как построить уравнение регрессии вида $y = F(x_1)$. Если исследователь изучает влияние нескольких переменных $x_1, x_2 \cdots x_k$ на результатирующий признак y, то возникает необходимость в умении строить регрессионное уравнение более общего вида, то есть $y = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k$ – уравнение множественной регрессии, где $a, b_1, b_2, ...b_k$ – постоянные коэффициенты, называемые частными коэффициентами регрессии.

В связи с последним уравнением необходимо рассмотреть следующие вопросы: а) как по эмпирическим данным вычислить коэффициенты регрессии $a, b_1, b_2, ...b_k$; б) какую интерпретацию можно приписать этим коэффициентам; в) как оценить тесноту связи между y и каждым из x_i в отдельности (при элиминировании действия остальных); г) оценить тесноту связи между y и всеми переменными $x_1, x_2 \cdots x_k$ в совокупности.

Рассмотрим этот вопрос на примере построения двухфакторного регрессионного уравнения. Предположим, что изучается зависимость недельного бюджета свободного времени (y) от уровня образования (x_1) и возраста (x_2) определенной группы трудящихся по данным выборочного исследования. Будем искать эту зависимость, обратившись к линейному уравнению следующего вида: $y = a + b_1 x_1 + b_2 x_2$.

При расчете коэффициентов уравнения множественной регрессии полезно преобразовать исходные эмпирические данные следующим образом:

$Z_{ij} = \dfrac{X_{ij} - \overline{X}_i}{\sigma_i}$, при этом уравнение множественной регрессии примет вид $y = c_1 z_1 + c_2 z_2$, где c_1 и c_2 находятся из следующей системы уравнений:

$$\begin{cases} c_1 + r_{12} c_2 = r_{1y} \\ c_1 r_{12} + c_2 = r_{2y} \end{cases},$$

решая которую, получаем, что $c_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}$, $c_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}$,

где r – коэффициент парной корреляции между признаками.

c_1 и c_2 называются стандартизированными коэффициентами регрессии. Следовательно, зная коэффициенты корреляции между изучаемыми признаками, можно подсчитать коэффициенты регрессии. Подставим конкретные значения всех возможных r из таблицы 4.21. Пусть в результате исследования N человек получены эмпирические значения, сведенные в следующую таблицу (в каждом столбце представлены несгруппированные данные):

Таблица 4.21

Матрица коэффициентов связи (теоретический пример)

	Y	X_1	X_2	Z_1	Z_2
1	y_1	x_1^1	x_1^2		
2	y_2	x_2^1	x_2^2		
3	y_3	x_3^1	x_3^2		
...	...	...	...		
n	y_n	x_n^1	x_n^2		
	$\bar{y}_1$	$\bar{x}_1$	$\bar{x}_2$		
	σ_y	σ_{x1}	σ_{x2}		

r_{12}	r_{1y}	r_{2y}
–0.027	0.556	–0.131

Тогда $c_1 = \frac{0.556 - (-0.131)(-0.027)}{1 - (-0.027)^2} = 0.55.$

Аналогично $c_2 =$ –0,12, и уравнение регрессии запишется в виде y=0,55z_1 – 0,12z_2.

Коэффициенты $a, b_1, b_2, ... b_k$ исходного регрессионного уравнения находятся по формулам:

$$b_1 = c_1 \frac{\sigma_y}{\sigma_1}; \qquad b_2 = c_2 \frac{\sigma_y}{\sigma_2};$$

$$a = \bar{y} - b_1 x_1 - b_2 x_2 .$$

Подставляя в эти формулы полученные данные, будем иметь:

$b_1 = c_1 \frac{\sigma_y}{\sigma_1} = 3{,}13;$ $\qquad$ $b_2 = c_2 \frac{\sigma_y}{\sigma_2} = 0{,}17;$

$a = \bar{y} - b_1 x_1 - b_2 x_2 = 8{,}56.$

Как же следует интерпретировать это уравнение? Например, значение b_2 показывает, что в среднем недельный бюджет свободного времени при увеличении возраста на один год и при фиксированном значении x_1 уменьшается на 0,17 часа. Аналогично интерпретируется b_1.

Коэффициенты b_1, b_2 можно в то же время рассматривать и как показатели тесноты связи между переменными y и, например x_1, при постоянстве x_2.

Аналогичную интерпретацию можно применять и к стандартизированным коэффициентам регрессии c_1 и c_2. Однако поскольку c_1 и c_2 вычисляются исходя из нормированных переменных, они являются безразмерными и позволяют сравнивать тесноту связи между переменными, измеряемыми в различных единицах. Например, в вышеприведенном примере x_1 измеряется в классах, а x_2 – в годах. c_1 и c_2 позволяют сравнить, насколько x_1 теснее связан с y, чем x_2.

Частный коэффициент корреляции – показатель, который характеризует тесноту и направление связи между результирующим признаком (y) и факторным признаком (x_i) при элиминировании остальных признаков.

Частный коэффициент корреляции записывается $r_{y1.2}$ и вычисляется по следующей формуле:

$$r_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}.$$

Для характеристики степени связи результирующего признака y с совокупностью независимых переменных служит множественный коэффициент корреляции $R^2_{y(1\ldots k)}$, который вычисляется по формуле (иногда он выражается в процентах):

$$1 - R^2_{y(1\ldots k)} = (1 - r^2_{y1})(1 - r^2_{y2.1})\ldots(1 - r^2_{yn.23\ldots(k-1)}).$$

Так, для вышеприведенного примера множественный коэффициент корреляции равен:

$$R^2_{y(12)} = 1 - (1 - r^2_{y1})(1 - r^2_{y2.1}) = 1 - (1 - 0{,}56^2)(1 - 0{,}140^2) = 0{,}323 = 32\%.$$

Множественный коэффициент корреляции показывает, что включение признаков x_1 и x_2 в уравнение y= 8,35+3,14x_1−0,166x_2 на 32% объясняет изменчивость результирующего фактора. Чем больше $R^2_{y(1\ldots k)}$, тем полнее независимые переменные $x_1, x_2 \cdots x_k$ описывают признак y. Обычно $R^2_{y(1\ldots k)}$ служит критерием включения или исключения новой переменной в регрессионное уравнение. Если $R^2_{y(1\ldots k)}$ мало изменяется при включении новой переменной в уравнение, то такая переменная отбрасывается.

Вопросы и задания для самоконтроля

1. *Дайте определения: взаимная сопряженность, совместное появление, ковариация.*

2. Охарактеризуйте следующие связи как простые или сложные и дайте свое обоснование: продолжительность брака и размер семьи; доход и социальный статус; смертный приговор и уменьшение случаев убийства; психическое расстройство и самоубийство; диаметр и длина окружности.

3. Дайте определения: корреляционное поле, прямолинейность, криволинейность, корреляционная таблица, маргинальные распределения, вариабельность, двумерная совокупность данных.

4. Если маргинальные распределения нормальны по форме, будет ли рассеяние точек обязательно гомоскедастичным?

5. Дайте определения: таблица 2×1, таблица 2×2, статистическая связь, односторонняя и двусторонняя связь, ожидаемая и наблюдаемая частота, маргинал, распределение совместных частот.

6. При исследовании психически ненормальных людей 94% продемонстрировали повышенную конфликтность в поведении перед началом психического заболевания: А). Доказывает ли это наблюдение связь между состоянием конфликта и психическим заболеванием, и почему Вы думаете именно так? Б). Какой бы Вы сделали вывод, если бы 94% из группы «нормальных» людей так же переживали состояние конфликта? В). Если бы 50% из группы «нормальных» людей переживали состояние конфликта, какой бы вывод вы сделали? Г). Достаточно ли иметь таблицу 2×1 для того, чтобы доказать наличие связи?

7. Дайте определения: ранг, ранговый порядок, равные интервалы, порядковые числа, корреляция ранговых последовательностей.

8. При каких обстоятельствах ранжируются качественные данные?

9. Приведите пример уменьшения корреляции в случае объединения рангов.

10. Опишите алгоритм вычисления коэффициента Спирмена.

11. Каков принцип вычисления множественного коэффициента

корреляции для ранговых рядов?

12. Дайте определения: коэффициент детерминации, объясняемая вариация, коэффициент недетерминированности, необъясняемая вариация, линия регрессия, линия наименьших квадратов, коэффициент корреляции моментов произведений, угол наклона, общее отклонение, необъясняемое отклонение.

13. Как следует понимать необъясняемые остатки: как результат случая либо как результат воздействия определяющих факторов?

14. Предположим, что r *=0,3 для связи между школьными оценками и часами подготовкик занятиям. Проанализируйте эту «низкую корреляцию». Действительно ли подготовка не влияет на оценки?*

15. Для использования r *требуется, чтобы рассеяние наблюдений относительно линии регрессии было гомоскедастичным. Объясните, почему?*

16. Допустимо ли вычисление η^2*, если линия регрессии плавная?*

17. Можно ли вычислить η^2 *для двух качественных переменных? Почему?*

18. Как порядок столбцов (строк) влияет на числовое значение η^2 *?*

19. При каких табличных условиях числовые значения r^2 *и* η^2 *совпадают?*

20. Покажите графически условия, при которых значение r^2 *приблизительно равнялось бы нулю, а* η^2 *– единице.*

21. Проверьте графически, что значение η^2 *можно сделать близким к единице, используя столько же столбцов, сколько имеется величин.*

22. Подсчитайте η^2_{yx} *и* η^2_{xy} *для динамики сети высших учебных заведений Украины III–IV уровня аккредитации и численности студентов в них (см. Таблица 4.13, с. 170).*

23. Что фиксирует частный коэффициент корреляции?