Глава II

КОРРЕЛЯЦИИ

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

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

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

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

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

[65]

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

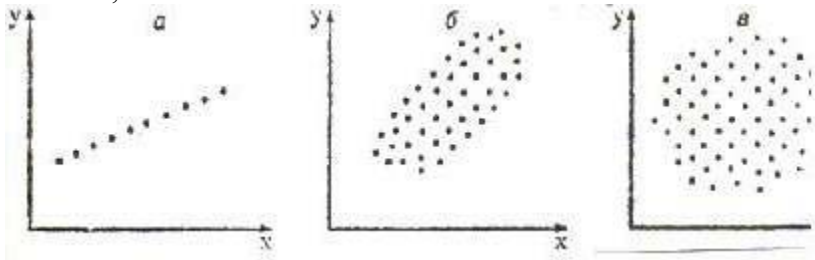

Рис. 19. Корреляционное поле для различных видов связи: а — функциональная связь; б—корреляционная связь; в— отсутствие связи.

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

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

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

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

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

[66]

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

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

$PV = CMT$

(здесь *C* — константа)

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

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

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

[67]

Пример 11. Для признаков *X* и *Y,* задаваемых таблицей 13, коэффициент корреляции (см. § 4 этой главы) $r = 0{,}98$, т.е. между *X* и *Y* есть значимая прямая связь. Здесь: *X* — общая

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

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

заработная плата школьных работников в миллионах долларов, а *Y* — общее потребление вина и ликеров в США в миллионах галлонов. Едва ли можно утверждать, что заработная плата школьных работников непосредственно зависит от потребления вина и ликеров или потребление винно-ликерных изделий от зарплаты школьных работни-

Таблица 13

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

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

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

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

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

[68]

причины. Связи такого рода иногда называют связями сопутствия.

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

Таблица 14

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

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

сравнивать влияние различных факторов (частная корреляция).

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

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

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

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

[69]

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

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

Таблица 15

Общий вид корреляционной таблицы двух признаков.

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

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

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

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

[70]

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

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

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

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

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

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

и т.д.

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

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

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

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

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

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

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

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

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

[71]

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

Рассмотрим клетку (i, j). Она находится в i-ой строке, на долю которой приходится $N(x_i)$ объектов. Если связи нет, то число объектов в данной клетке будет определяться только общим числом объектов в столбце: чем больше $N(y_j)$, тем больше их окажется и в клетке (i, j), т.е. на ее долю придется $\frac{1}{N} \bullet N(y_j)$ частей $N(x_i)$. Итак, если связи нет, то в (i, j) попадет

$\frac{1}{N}N(x_i)\bullet N(y_j)$ объектов. Станем обозначать эту частоту N_{ij}^0 и называть теоретической в отличие от фактически наблюдаемой — эмпирической $N_{ij} : N_{ij}^0 = \frac{1}{N}N(x_i)N(y_j)$.

Какова мера отклонения эмпирической таблицы от теоретической?

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

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

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

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

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

[72]

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

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

Таблица 16

Зависимость между возрастом и отношением к моде

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

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

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

Нам заданы $N(x_i)$ и $N(y_j)$. Характер связи X с Y определит распределение объектов по $k \times l$ клеткам таблицы. Так как сумма частот клеток строки (как и столбца) фиксирована, то на распределение объектов по клеткам в каждой строке и в каждом столбце наложено по одному ограничению. Общее число ограничений $k+l$ должно быть уменьшено на 1, так как

[5] О понятии вероятности см. Приложение 1.

эти ограничения не независимы: сумма итогов столбцов равна сумме итогов строк (и равна *N).* Следовательно, на распределение объектов по *k•l* клеткам таблицы наложено *k+l—1* ограничение. Величина $f = kl - (k + l - 1) = (k - 1)(l - 1)$ называется числом степеней свободы корреляционной таблицы.

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

[73]

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

Таблица 17

Пример расчета χ^2

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

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

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

Рассмотрим эмпирическую корреляционную таблицу 17.

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

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

[74]

[6] Приложение 3, таблица Б (χ_0^2).

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

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

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

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

Вопрос о мере связи будет рассмотрен позднее.

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

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

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

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

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

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

Определить, есть ли связь между национальностью мужа и жены.

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

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

Таблица 18

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

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

Таблица 19

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

X	Y			
	y_1	y_2	y_3	$N(x_i)$
x_1	689	30	37	756
x_2	758	53	91	902

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

x_3	76	3	4	83
$N(y_j)$	1523	86	132	1741

Таблица 20

Связь между национальностями мужа и жены

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

[76]

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

Таблица 21

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

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

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

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

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

[77]

ответить на вопрос, соответствует ли эмпирическое распределение теоретическому:

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

где N_i *и* N_i^0 — эмпирическая и теоретическая частоты, а n — число вариантов. Формулу (II,1,1) можно рассматривать как частный случай формулы (II,1,3) для распределения с числом вариант $n=k•l$. Теоретические частоты могут определяться на основании некоторой

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

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

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

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

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

[78]

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

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

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

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

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

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

[79]

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

Таблица 22

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

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

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

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

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

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

[80]

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

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

Если связь функциональная (т.е. каждому *x* соответствует одно вполне определенное значение *y*), то без ограничения общности можно считать, что корреляционная таблица должна иметь диагональный вид. Пусть для определенности *k<l,* тогда

$N_{ij} = \left\{ \begin{matrix} 0, i \neq j \\ N_{ij}, i = j = \overline{1,k} \end{matrix} \right\}$ и так как $N(x_i)=N(y_j)$, то $N_{ij}^0 = N_{ij}^2 / N$. Теперь просто найти $\chi^2_{\max}$. Подставляя N_{ij}^0 в (II,2,1) получаем: $\chi^2_{max} = N(k-1)$. При $k>l$ аналогично $\chi^2_{max} = N(l-1)$. Таким образом,

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

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

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

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

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

Пирсон предложил рассматривать величину

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

которая получила название *коэффициента средней квадратической сопряженности Пирсона.*

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

[81]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[83]

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

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

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

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

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

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

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

[84]

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

Таблица 23

Общий вид таблицы 2 × 2

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

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

$N(y_j)$	$N(y_1)$	$N(y_2)$	N

щими доказательства. Однако с подобными суждениями нельзя согласиться.

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

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

[85]

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

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

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

Таблица 24

Зависимость между производительностью труда и удовлетворенностью профессией

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

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

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

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

[86]

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

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

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

19	1		15	5		10	10
31	49		35	45		40	40
а			б			в	

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

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

Таблица 25

Зависимость между учебой и участием в рационализации

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

[87]

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

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

Коэффициент контингенции Ф по определению:

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

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

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

убедиться на примерах). У разных таблиц разные Φ_{max}, поэтому Φ, рассчитанные для них, часто несопоставимы.

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

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

[88]

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

Итак, по данным на 1939 г.:

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

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

По данным за 1959 г.:

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

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

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

Для мужчин соответствующая таблица (1939 г.):

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

[89]

Для женщин:

X	Y		$N(x_i)$
	y_1	y_2	
x_1	5,18	26,31	31,49
x_2	1,83	65,95	67,78

$N(y_j)$	7,01	92,26	99,27
Q=0,75			

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

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

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

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

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

[90]

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

Составить таблицу, вычислить Q.

Ответ: Q=0,44.

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

Ответ: Q=0,52.

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

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

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

Сделаем одно очень существенное замечание. Изучаемые социологами совокупности часто оказываются весьма разнородными. Например, рабочие предприятия —люди разных профессий, разного пола, возраста, образования и т.д. При достаточно разнородной

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

совокупности могут возникать кажущиеся связи, либо оказаться скрытыми действительные. Поясним это примером.

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

[91]

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

Для нее Q, очевидно, равно нулю.

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

а

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

б

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

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

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

Этот пример формально иллюстрирует случай, когда связь оказалась скрытой.

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

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

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

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

[92]

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

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

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

4. Коэффициент ранговой корреляции Спирмена

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

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

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

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

[93]

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

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

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

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

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

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

При $N=2k$:

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

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

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

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

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

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

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

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

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

будет обладать требуемым свойством.

[94]

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

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

Таблица 26

Вычисление коэффициента Спирмена ρ

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

фициентом текучести, значения этих коэффициентов находятся в третьей колонке

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

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

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

15-20	0,65	3,3
Свыше 20	0,73	1,3

Вычислить ρ.

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

[95]

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

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

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

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

где

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

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

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

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

[96]

5. Коэффициент парной корреляции и его связь с другими коэффициентами

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

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

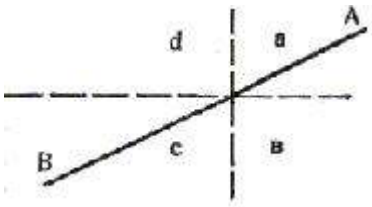

Рис. 20. Области корреляционного поля

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

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

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

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

[97]

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

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

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

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

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

Перейдем к графической интерпретации.

На рис. 20 прямые $y = \overline{y}$ и $x = \overline{x}$ разбивают координатную плоскость на 4 части: *a, b, c, d.* Положительность *S* означает преимущественное расположение точек корреляционного

поля в областях *a* и *c* (отрицательность — *b* и *d).* Величина *S* близка к нулю, если поле равномерно «размазано».

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

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

[98]

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

Окончательно:

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

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

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

если использовать определение среднего квадратического отклонения.

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

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

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

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

1. Пусть *X* и *Y* принимают такие значения:

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

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

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

2. Пусть *X* и *Y* принимают значения:

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

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

[99]

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

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

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

Ответ: *r*=0

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

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

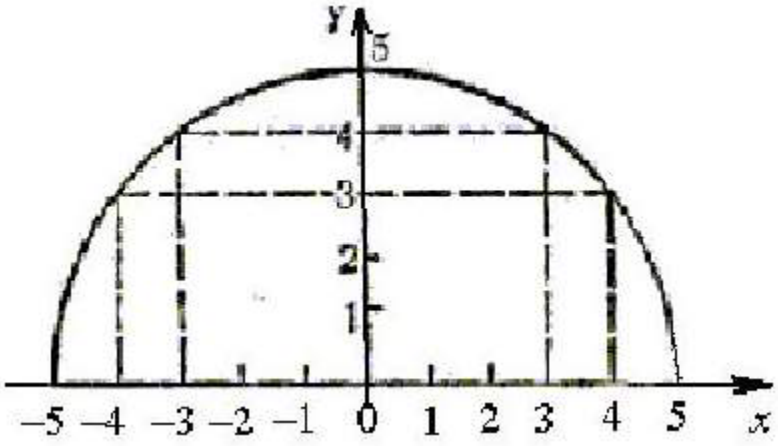

Рис.21. Криволинейная функциональная зависимость

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

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

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

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

Представим эту зависимость графически (рис. 21).

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

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

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

[100]

принимает вид:

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

Указание: использовать (II,5,3).

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

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

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

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

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

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

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

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

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

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

[101]

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

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

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

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

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

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

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

[102]

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

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

(см. Приложение 2).

Аналогично

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

В обозначениях предыдущего параграфа:

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

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

Отсюда

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

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

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

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

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

[103]

следовательно, теперь

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

Итак,

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

Завершим здесь рассмотрение р выводом формулы для случая объединенных рангов.

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

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

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

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

Здесь мы воспользовались формулами Приложения 2.

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

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

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

если объединить ранги X.

Аналогичный вклад T_y дает объединение рангов по У:

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

[104]

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

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

Теперь

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

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

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

Таким образом,

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

В заключение параграфа приведем пример вычисления ρ с объединением рангов.

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

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

Имеем $T_x = 10{,}5$; $T_y = 1$;

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

[105]

книгах использовалась следующая формула для расчета ρ:

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

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

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

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

Таблица 27

Пример вычисления коэффициента Спирмена ρ с объединением рангов

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

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

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

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

[106]

6. Коэффициент ранговой корреляции Кендэла

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

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

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

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

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

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

Величина

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

называется коэффициентом Кендэла.

[107]

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

2. Показать, что

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

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

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

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

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

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

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

Упорядочим ранги по *X:*

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

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

[108]

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

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

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

Вычислить коэффициент Кендэла.

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

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

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

Ответ: τ = – 0,24

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

Возрастная группа	K_T (%)	J_p	ранг по X(K_T)	ранг по Y(J_p)

до 18 лет	12,9	0,57	5	5
18–19	13,0	0,38	4	7
20–21	17,1	0,35	3	8
22–24	37,1	0,24	1	9
25–30	19,9	0,39	2	6
31–40	7,9	0,59	6	4
41–50	5,6	0,69	9	3
51–60	6,1	0,76	8	2
свыше 60 лет	6,4	0,77	7	1

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

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

P=5 Q=31

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

[109]

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

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

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

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

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

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

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

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

[110]

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

$(\sum_s n_S = 24)$

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

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

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

Пусть $N = 3$, имеем 6 перестановок:

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

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

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

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

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

[111]

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

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

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

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

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

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

$\sigma^2 = (1/18)N(N—1\ )(2N+5)$

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

[112]

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

Введем величину

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

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

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

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

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

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

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

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

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

Аналогично $\tau_{min} = -1$.

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

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

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

Числитель можно несколько упростить.

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

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

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

Итак, мы познакомились с новой формой записи коэффициента Кендэла (II,6,4).

Далее, допустим, что *t* рангов по X с *l*+ 1 по $l + t$ объединены, т.е. ранговый ряд имеет вид:

$$1,2,...,l,l+\frac{1+t}{2},l+\frac{1+t}{2},...l+\frac{1+t}{2},l+t+1,...,N$$

Сопоставление всех не объединенных рангов с другими, объединенными и не объединенными, дадут те же результаты, что и ранее: в нашем примере ранг объединенных все равно выше рангов 1, 2, ..., *l* и ниже рангов *l*+ *t*+ 1, ..., *N*. Но сопоставление объединенных рангов между собой не будет порождать ни +1, ни –1, так как эти ранги равны. Доопределим теперь a_{rs} и

b_{rs}, так, чтобы $a_{rs} = b_{rs} = 0$ при совпадении рангов (это естественно). Всего сопоставлений объединенных рангов $\frac{t(t-1)}{2}$. Сумма $\sum_r \sum_s a_{rs}^2$ уменьшится на $\frac{t(t-1)}{2}$. Если объединений несколько, скажем, p, а t_v – число объединенных рангов в *v*-ом объединении по X,

[114]

то сумма уменьшится на величину

$$U_x = \sum_{v=1}^{p} \frac{t_v(t_v-1)}{2}$$

Пусть q – число объединенных рангов y, а u_w – число объединенных рангов в *w*-ом объединении, тогда сумма $\sum\sum b_{rs}^2$ уменьшится на

$$U_y = \sum_{w=1}^{q} \frac{u_w(u_w-1)}{2}$$

Итак, для случая объединенных рангов окончательно имеем:

$$\tau = \frac{P-Q}{\sqrt{(\frac{N(N-1)}{2}-U_x)(\frac{N(N-1)}{2}-U_y)}}. \text{ (II,6,5,)}$$

В отличие от ρ коэффициент τ без поправки меньше, чем коэффициент τ с поправкой, т.е. использование τ без поправок повышает ошибку II рода и менее опасно, чем использование ρ без поправок (см. гл. V).

Пример 22. Рассмотрим следующую таблицу:

Объекты	A	B	C	D	E	F	G	H	K	L	M	N
X	1,5	1,5	3	4	6	6	6	8	9,5	9,5	11	12
Y	2,5	2,5	7	4,5	1	4,5	6	11,5	11,5	8,5	8,5	10

Что порождает в S элемент A?

При сопоставлении A с S, очевидно, 0 (одинаковые ранги по X), A с C – плюс единицу (+1) × (+1) = 1, аналогично 1 порождает сопоставление A с D, F, G, H, K, L, M, N; при сопоставлении A с E появляется минус единица (ранг по X в прямой, а по Y – в обратной последовательности: 1 × (–1) = –1).

Таким образом, вклад A в S равен +8. Продолжая сопоставления, получим: S = 8 + 8 + 1 +5 + 5 + 5 + 5 – 3 – 2 + 1 + 1 = 34.

В X – последовательности три объединения: $t_1 = 2$; $t_2 = 3$; $t_3 = 2$; $U_x = 5$; во второй – четыре: $u_1 = u_2 = u_3 = u_4 = 2$; $U_y = 4$. Теперь по формуле (II,6,5): τ = 0,55.

Упражнение 53. В упоминавшейся книге «Методика и техника статистической обработки первичной социологической информации» приводится таблица «Вычисление

[115]

коэффициента корреляции рангов Кендэла между ответами рабочих: «интересная работа» и «образование соответствует работе» (с. 17). Воспроизведем часть ее.

Рассчитать τ. В случае необходимости помочь в этом может цитируемая книга. Там, в частности, показывается, что

Таблица 28

Пример вычисления коэффициента ранговой корреляции Кендэла

Номер профессиональной группы	X –, ответившие, что работа интересная, %	ранг по X	У – лица, ответившие, что образование соответствует работе, %	ранг по У

1	100,0	3	100	1
2	100,0	3	87,5	5,5
3	100,0	3	77,0	9
4	100,0	3	75,0	10
5	100,0	3	50,0	11,5
6	83,5	6,5	92,0	3
7	83,5	6,5	83,5	8
8	83,0	8	90,0	4
9	82,5	9	94,5	2
10	71,0	10	87,0	7
11	55,5	11	87,5	5,5
12	50,0	12	50,0	11,5
13	28,5	13	43,0	13
14	0	14	0	14

P = 61, Q = 28, однако при вычислении τ не учтено, что имеются объединения рангов. Даже если Вы используете книгу, рассчитайте τ самостоятельно, с учетом объединений. Для контроля: $U_x = 1$, $U_y = 2$. Ответ: τ = + 0,39.

Об оценке существенности τ в случае объединенных рангов см. § 8 главы V.

До сих пор использовались формулы, справедливые для любых *N*, однако удобные лишь для малых (не более 20–30); в противном случае вычисления существенно затрудняются.

Сейчас мы рассмотрим большие *N*. В таких случаях признаки шкалируются. Как и ранее, будем считать, что признак X принимает значения x_i где $i = \overline{1,k}$, а признак Y – значения y_j, где $j = \overline{1,l}$ (обычно *k*, *l* ≈ 5–10). Эмпирический материал сводится в корреляционную таблицу $\{N_{ij}\}$, для которой $\sum_i \sum N_{ij} = N$ (см. § 1, главы II).

[116]

В качестве исходной возьмем формулу

$$\tau = \frac{S}{\sqrt{A \cdot B}},\ A = \sum_r \sum_s a_{rs}^2,\ B = \sum_r \sum_s b_{rs}^2$$

$$S = \sum_r \sum_s a_{rs} b_{rs}\ . \qquad \text{(II,6,6)}$$

При больших *N* выполнить суммирование по r и s от 1 до *N* чрезвычайно затруднительно, поэтому перейдем к суммированию по *i* и *j* от 1 до *k* и *l* соответственно.

Рассмотрим A. Нам нужно сравнить ранги по X каждой пары объектов, а результаты просуммировать[22]. Очевидно, можно не сравнивать между собой элементы строки, так как у них одинаковые ранги по X. Следовательно, все элементы, у которых $X = x_1$ (всего их $N(x_1)$), можно не сравнивать друг с другом, но следует сравнить с элементами, у которых $X = x_2$. Такое

[22] В дальнейшем изложении предполагается, что значения X и Y выписаны в таблице в порядке возрастания (сверху вниз и слева направо).

сравнение породит $N(x_1) \cdot N(x_2)$ единиц, а сравнение элементов с $X = x_1$ с элементами, у которых $X = x_3$, дает $N(x_1)\,N(x_3)$ единиц и т.д. Поэтому

$$A = N(x_1)[N(x_2) + N(x_3) + \ldots + N(x_k)] + N(x_2)[N(x_3) + N(x_4) + \ldots + N(x_k)] + \ldots + N(x_{k-1})N(x_k) =$$

$$A = N(x_1)\left[N(x_2) + N(x_3) + \ldots + N(x_k)\right] + N(x_2)\left[N(x_3) + N(x_4) + \ldots + N(x_k)\right] + \ldots +$$

$$+ N(x_{k-1})N(x_k) = \sum_{i=1}^{k-1} N(x_i) \sum_{p=1}^{k-1} N(x_{i+p}) \qquad \text{(II,6,7)}$$

Упражнение 54. Показать, что

$$B = \sum_{j=1}^{l-1} N(y_j) \sum_{q=1}^{l-j} N(y_{j+q}) \qquad \text{(II,6,8)}$$

Перейдем к рассмотрению S. Теперь для каждой пары элементов нужно сравнивать и ранги по X (a_{rs}), и ранги по Y (b_{rs}).

Рассмотрим элементы клетки (i, j). Ясно, что их не нужно сравнивать ни с элементами i-ой строки (об этом мы уже говорили), ни с элементами j-го столбца (у элементов столбца одинаковые ранги по Y, следовательно, за счет b_{rs} соответствующее слагаемое обратится в нуль). Станем сравнивать некоторый элемент из клетки (i, j) с элементом клетки (i', j'), если $i'>i$, $j'> j$. Такое сравнение для каждой пары объектов породит +1 в силу упорядоченности пунктов шкалы ($a_{rs} = 1$, $b_{rs} = 1$). Если $i'>i$, а $j'> j$, то каждая пара породит –1 ($a_{rs} = 1$, $b_{rs} = -1$). Суммируя по i',j', мы

[117]

переберем всевозможные сравнения выделенного элемента из клетки (i, j) со всеми элементами, лежащими ниже и справа ($j'> j$, $i'>i$) которые дадут, таким образом, $\sum_{i'=i+1}^{k} \sum_{j'=j+1}^{l} N_{i'j'}$. Сопоставление элемента из клетки (i,j) с элементами, расположенными ниже и слева от этой клетки, порождает слагаемое $\sum_{i'=i+1}^{k} \sum_{j'=1}^{j-1} N_{i'j'}$. Так как все элементы клетки (i,j) равно-

Таблица 29

Связь удовлетворенности работой с удовлетворенностью специальностью

X	Y			$N(x_i)$
	удовлетворен	промежуточная позиция	не удовлетворен	
удовлетворен	1472	50	65	1587
промежуточная позиция	136	65	42	243
не удовлетворен	126	42	165	333
$N(y_j)$	1734	157	272	2163

правны, то умножая результат на N_{ij} и суммируя затем по i и j, мы осуществим вообще все возможные сравнения пар элементов.

Упражнение 55. Почему не нужно рассматривать случай $i'<i$?

Итак,

$$S = \sum_{i=1}^{k} \sum_{j=1}^{l} N_{ij} \left(\sum_{i'=i+1}^{k} \sum_{j'=j+1}^{l} Ni'j' - \sum_{i'=i+1}^{k} \sum_{j'=1}^{j-1} Ni'j' \right) \quad \text{(II,6,9)}$$

Тем самым мы завершили переход к корреляционной таблице во всех множителях τ[23].

Для иллюстрации этой «страшной» формулы приведем пример, который покажет справедливость пословицы «не так страшен черт, как его рисуют».

[23] Авторы выражают благодарность Г.И. Саганенко за помощь при выводе соотношения (II,6,9).

Пример 23. Изучая связь удовлетворенности работой (Y) с удовлетворенностью специальностью (X) мы, в частности, получили корреляционную таблицу 29 (массив, ОСРЗ).

[118]

Теперь A = 1587 (243 + 333) + 243·333 = 995031;

B = 1734 (157 + 272) + 157·272 = 786590;

S = 1472 (65 + 42 + 42 + 165) + 50 (42 + 165 – 136 – 126) – 65 (136 + 65 + 126 + 42) + 136 (42 + 165) + 65 (165 – 126) – 42 (126 + 42) = 459104;

$\tau = +0{,}52$.

Таким образом, между изучаемыми удовлетворенностями есть тесная положительная связь.

Упражнение 56. Для признаков удовлетворенность работой (Y), удовлетворенность общественной работой (X) корреляционная таблица имеет вид:

Таблица 30

Связь удовлетворенности работой (Y) с удовлетворенностью общественной работой (X)

X	Y			N(x_i)
	Y_1	Y_2	Y_3	
x_1	1241	82	150	1473
x_2	147	11	38	196
x_3	103	13	13	129
N(yj)	1491	106	201	1798

Вычислить τ. Ответ: $\tau = +0{,}31$.

Связь, таким образом, тоже положительная, но менее тесная. Еще менее тесной, например, оказывается связь между удовлетворенностью работой и удовлетворенностью досугом (для соответствующей корреляционной таблицы $\tau = +0{,}14$), что допускает естественную интерпретацию.

Коэффициент τ, определяемый формулой (II,6,6), может обращаться в ±1 только в том случае, когда таблица диагональна.

В самом деле, согласно неравенству Коши[24] |S| максимален, если наборы a_{rs} и b_{rs} пропорциональны: $b_{rs} = \alpha \cdot a_{rs}$. Это возможно лишь тогда, когда все наблюдения либо на положительной ($\alpha = 1$), либо на отрицательной ($\alpha = -1$) главной диагонали таблицы, т.е. если таблица квадратная (если есть не диагональные элементы, то α не будет знако-

[119]

постоянной величиной, соотношение $b_{rs} = \alpha\, a_{rs}$ не будет выполняться для всех пар элементов).

Для прямоугольной таблицы |S| достигает максимума, если: 1) все наблюдения лежат в клетках самой длинной диагонали таблицы, т.е. диагонали, содержащей $m = min\,(k, l)$ клеток, так как в случае появления недиагональных элементов в S, кроме нулей типа 0·0, добавляются нули типа $a_{rs} \cdot 0$ и $0 \cdot b_{rs}$, причем за счет уменьшения числа слагаемых, равных 1;

2) все наблюдения равномерно распределены между диагональными клетками, т.е. $N_{ii} = N/m$ (так как обычно $N >> m$, то можно считать, что оно кратно m без существенной потери точности).

Проиллюстрируем первое утверждение, например, для следующей таблицы:

[24] Для читателя, незнакомого с этим неравенством, мы приводим его вывод в конце параграфа.

X	Y		N(x_i)
	y_1	y_2	
x_1	N_{11}	1	$N_{11}+1$
x_2	0	$N_{22}-1$	$N_{22}-1$
x_3	0	0	0
N(y_j)	N_{11}	N_{22}	N

$$S = N_{11}(N_{22}-1) \prec N_{11}N_{22}$$

Проиллюстрируем второе утверждение. Рассмотрим, например, диагональную таблицу 3 × 3:

X	Y			N(x_i)
	y_1	y_2	y_3	
x_1	N_{11}	0	0	N_{11}
x_2	0	N_{22}	0	N_{22}
x_3	0	0	N_{33}	N_{33}
N(y_j)	N_{11}	N_{22}	N_{33}	N

Для нее

$$S = N_{11}N_{22} + N_{11}N_{33} + N_{22}N_{33} \le N_{11}^2 + N_{22}^2 + N_{33}^2,$$

$$S_{\max} = N^2/3 при N_{11} = N_{22} = N_{33} = N/3,$$

т.е.

[120]

если все наблюдения распределены равномерно. Здесь мы использовали известное неравенство

$$ab + bc + ac \le a^2 + b^2 + c^2,$$

которое легко получить, складывая почленно три очевидных неравенства

$$(a-b)^2 \ge 0, (a-c)^2 \ge 0, (b-c)^2 \ge 0.$$

В общем случае в каждой клетке самой длинной диагонали должно быть N/m элементов. Сопоставляя элементы первой клетки с остальными, мы получим $\frac{N}{m}\cdot\frac{N}{m}(m-1)$ единиц, а элементы второй с прочими $\frac{N}{m}\cdot\frac{N}{m}(m-2)$, так как их уже не нужно сравнивать с элементами первой и т.д.

В итоге

$$S_{\max} = \frac{N^2}{m^2}\left[(m-1)+(m-2)+\cdots+2+1\right] = \frac{N^2(m-1)}{2m}$$

Но при этом значении S коэффициент τ, вообще говоря, не достигает значений ± 1.

Введем

$$\tau_c = \frac{S}{S_{\max}} = \frac{2mS}{N^2(m-1)}. \qquad \text{(II,6,10)}$$

Очевидно, он принимает значения, которые могут достичь ± 1 (если не считать незначительного эффекта, возникающего в случае, когда N не кратно *m)* даже для прямоугольных таблиц.

Коэффициент, определяемый (II,6,6), обозначают иногда τ_b, а (II,6,1) – τ_a, если нет объединений рангов $\tau_a = \tau_b$.

Обратим внимание на то, что три коэффициента r, ρ, τ можно рассмотреть с единой точки зрения. Действительно, пусть, как обычно, имеется совокупность из *N* индивидов, каждый из которых может быть охарактеризован с помощью значений двух признаков *X* и *Y*.

Выберем пару индивидов, например, *i* и *j* и станем приписывать ей некоторую *x* – оценку a_{ij} (конкретизация оценок будет дана ниже), обладающую свойством антисимметричности: $a_{ij} = -a_{ij}$ Аналогично введем *y* – оценку b_{ij}.

[121]

Рассмотрим величину

$$\Gamma = \frac{\sum_i \sum_j a_{ij} b_{ij}}{\sqrt{\sum_i \sum_j a_{ij}^2 \cdot \sum_i \sum_j b_{ij}^2}}$$

Мы уже видели (II,6,6), что для величины

$$a_{ij} \begin{cases} 1, если R_i^{(x)} \succ R_j^{(x)} \\ -1, если R_i^{(x)} \prec R_j^{(x)} \end{cases}$$

(где $R_i^{(x)}$ – ранг по X *i*-го элемента) и аналогичной величины b_{ij}: Γ = τ.

Пусть

$a_{ij} = x_j - x_i$, а, $b_{ij} = y_j - y_i$

тогда

$$\sum_i \sum_j (x_j - x_i)(y_j - y_i) = 2N \sum_i x_i y_i - 2 \sum_i \sum_j x_i y_j$$

$$\sum_i \sum_j (x_j - x_i)^2 = 2N \sum_i x_i^2 - 2(\sum_i x_i)^2$$

Теперь

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

Если положить $a_{ij} = R_j^{(x)} - R_i^{(x)}$, а $b_{ij} = R_j^{(y)} - R_i^{(y)}$, то можно аналогично предыдущему показать, что Γ обращается при этом в ρ. Это рассмотрение составит для читателя самостоятельное *упражнение 57*.

Мы же сошлемся на § 5 главы II, где было показано, что ρ является r, примененным к рангам, а так как для r рассмотрение проведено, то с точки зрения строгости изложения, выкладки данного упражнения в тексте книги не являются необходимыми. В заключение выведем неравенство Коши.

Очевидное неравенство $(A_{ij} - B_{ij})^2 \geq 0$ можно переписать в виде $\frac{1}{2} A_{ij}^2 + \frac{1}{2} B_{ij}^2 \geq A_{ij} B_{ij}$

Полагая

$$A_{ij} = \frac{a_{ij}}{\sqrt{\sum \sum a_{ij}^2}} \text{ и } B_{ij} = \frac{b_{ij}}{\sqrt{\sum \sum b_{ij}^2}}$$

[122]

и суммируя всевозможные неравенства, получим:

$$\frac{1}{2}\frac{\sum_i\sum_j a_{ij}^2}{\sum_i\sum_j a_{ij}^2}+\frac{1}{2}\frac{\sum_i\sum_j b_{ij}^2}{\sum_i\sum_j b_{ij}^2}\geq\frac{\sum_i\sum_j a_{ij}b_{ij}}{\sqrt{\sum_i\sum_j a_{ij}^2\sum_i\sum_j b_{ij}^2}}$$

Так как левая часть равна 1, то неравенство Коши доказано. Нетрудно видеть, что оно превращается в равенство, если все $a_{ij} = \alpha b_{ij}$ (убедиться подстановкой!), что и было нами ранее использовано.

Наконец, рассмотрим случай, когда оба признака измерены на уровне наличия – отсутствия.

Пусть индекс 1 соответствует наличию, а 2 отсутствию признака, тогда корреляционная таблица для признаков X и Y принимает вид:

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

Каждый элемент первой клетки положительной диагонали при сопоставлении с элементом второй породит +1, всего таких +1 в S войдет $N_{11}\cdot N_{22}$.

Сравнение элементов отрицательной диагонали породит $N_{12}\cdot N_{21}$, отрицательных единиц.

Следовательно,

$$S = N_{11}N_{22} - N_{12}N_{21};$$

$$U_x = \frac{1}{2}N(x_1)\left[N(x_1)-1\right]+\frac{1}{2}N(x_2)\left[N(x_2)-1\right];$$ а

$$\frac{1}{2}N(N-1)-U_x = N(x_1)N(x_2)$$

Аналогично:

$$\frac{1}{2}N(N-1)-U_y = N(y_1)N(y_2)$$

[123]

теперь коэффициент Кендэла, определяемый (II,6,5):

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

таким образом, совпадает с коэффициентом Ф (II,3,2).

Этот результат проясняет смысл формально введенного ранее коэффициента контингенции.

7. Энтропийные меры в социологическом анализе

Пусть некоторое событие может иметь k различных исходов $A_i(i=\overline{1,k})$ вероятность которых обозначим через $P(A_i)$. Ясно, что $\sum_{i=1}^{k}P(A_i)=1$. Например, при подбрасывании симметричной монеты $k = 2$, A_1 — выпадение герба, A_2 — решки, $P(A_1)=P(A_2)=\frac{1}{2}$

Допустим, что мы хотим предсказать исход испытания. Если k = 1, то исход предопределен. Если k = 2, то появляется неопределенность, которая максимальна при $P(A_1)$ =

= $P(A_2)$. Если $P(A_1) > P(A_2)$, то чем больше $P(A_1)$, тем меньше неопределенность предсказания. В пределе, когда $P(A_1) = 1$ ($P(A_2) = 0$), неопределенность исчезает: во всех испытаниях осуществляется исход A_1.

Чем больше k, тем менее определенны предсказания, тем больше неопределенность. По К. Шеннону, мерой неопределенности является величина $E = -\sum_{i=1}^{k} P(A_i)\log P(A_i)$, называемая *энтропией*. Если неопределенности нет и, скажем, реализуется l-ое состояние, т.е. $P(A_l) = 1$, а все остальные $P(A_i) = 0$, то E очевидно, обращается в нуль. Неопределенность максимальна, если все исходы равновозможны, т.е. $P(A_i) = 1/k$. При этом $E_{\max} = \log k$. Чем больше k, тем больше $E_{\max}$. Итак, $0 \le E \le \log k$.

Пусть N индивидов некоторой совокупности обладают некоторым признаком X, и событие A_i состоит в том, что значение признака равно x_i. Обозначим через N_i число индивидов, у которых $X = x_i$. Если N достаточно велико, то $P_i = N_i/N$, а E – мера «распыленности» распределения. Для сопоставления различных распределений целесообразно перейти к нормированному коэффициенту $\varepsilon = E/E_{max}$. Величина ε, принимающая значения между 0 и 1, является *аналогом дисперсии*.

[124]

Перейдем к двухмерным распределениям для признаков X и Y в случае, когда эмпирический материал сведен в корреляционную таблицу $\{N_{ij}\}$.

Теперь

$$E = -\sum_{i=1}^{k}\sum_{j=1}^{l} P_{ij}\log P_{ij},$$

где $P_{ij} = N_{ij}/N$ и суммирование ведется по всем клеткам корреляционной таблицы. Здесь и далее мы не указываем основание логарифма, так как обсуждаемые относительные показатели ε и λ, от него не зависят.

Упражнение 58. Показать, что $E_{max} = \log kl$

Упражнение 59. Показать, что теперь

$$\varepsilon = \frac{N\log N - \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}\log N_{ij}}{N\log kl}$$

Это выражение используется для расчета *энтропийной меры дисперсии*

Рассмотрим теперь так называемую *энтропийную меру связи*. Неопределенность Y-распределения

$E_y = -\sum_{j=1}^{l}\frac{N(y_j)}{N}\log\frac{N(y_j)}{N}$, если ничего не известно об X-распределении.

Неопределенность Y-распределения у индивидов с $X = x_i$, так называемая условная неопределенность

$$E_{y/x_i} = -\sum_{j=1}^{l}\frac{N_{ij}}{N(x_i)}\log\frac{N_{ij}}{N(x_i)} \quad (i = \overline{1,k})$$

В итоговую условную неопределенность каждая строчка таблицы дает вклад с удельным весом $N(x_i)/N$, т.е. полная условная неопределенность Y-распределения:

$$E_{y/x} = \sum_{i=1}^{k}\frac{N(x_i)}{N}E_{y/x_i}$$

Мерой связи между признаками X и Y может служить величина относительной неопределенности

$$\lambda_{y/x} = \frac{E_y - E_{y/x}}{E_y}.$$

[125]

Упражнение 60. Рассмотреть для простейших таблиц 2x2 случай отсутствия связи и показать, что $\lambda = 0$. *Указание*: использовать, что $N_{ij} = N(x_i)N(y_j)/N$.

Упражнение 61. Рассмотреть случаи функциональной связи и показать, что $\lambda = 1$. *Указание*: учесть, что таблица принимает диагональный вид. Итак, $0 \le \lambda \le 1$. Чем больше λ, тем больше связь между признаками.

Упражнение 62. Вычислить ε и $\lambda_{y/x}$ для следующей таблицы:

X	Y						$N(x_i)$
	y_1	y_2	y_3	y_4	y_5	y_6	
x_1		14	28	48	66	45	202
x_2	1	35	53	40	36	8	173
x_3	3	39	15	13	4	2	76
$N(y_j)$	5	88	96	101	106	55	451

Ответ: $\varepsilon = 0{,}872$, $Ey/x_1 = 0{,}664$, $Ey/x_2 = 0{,}660$, $Ey/x_3 = 0{,}582$, $\lambda_{y/x} = 0{,}086$.

Упражнение 63. Обратимся к рассмотрению связи между удовлетворенностью работой и удовлетворенностью заработной платой. Для таблицы 18 (работники в возрасте до 30 лет) найти λ.

Ответ: $Ey/x_1 = 0{,}289$, $Ey/x_2 = 0{,}396$, $Ey/x_3 = 0{,}383$, $Ey/x = 0{,}348$, $\lambda = 0{,}030$.

Упражнение 64. Для таблицы 19 (работники старше 30 лет) найти λ. Ответ: $\lambda = 0{,}014$.

Таким образом, связь между рассматриваемыми показателями более тесная для молодых работников. В дальнейшем мы вернемся к этому вопросу еще раз, используя другие методы статистического изучения связей (§ 8 главы II).

Пример 24. Представляет несомненный интерес задача о связи интегральной удовлетворенности с частными удовлетворенностями (отдельными элементами рабочей ситуации).

В качестве элементов обычно выделяют: 1) содержание труда (совокупность трудовых функций, выполняемых в процессе создания потребительных стоимостей в процессе труда), 2) условия (факторы, под воздействием которых осуществляется трудовая деятельность: сменность, физическая нагрузка, состояние окружающей среды и т.д.);

[126]

3) организация (совокупность мероприятий, обеспечивающих рациональное использование рабочей силы); 4) оплата; 5) межличностные отношения и т.д.

Осознавая, что человек не может точно определить вклад, который вносит в общее состояние удовлетворенности удовлетворение отдельных потребностей, мы отказались от метода ранжирования различных факторов. Для изучения обсуждаемой связи использовались различные статистические показатели, которые вычислялись для распределений совокупности в случае, когда одним из признаков является интегральная удовлетворенность и другим – последовательно-частные.

Для T и λ элементы расположились так: содержание труда, организация, оплата, отношения с администрацией и т.д. (см. также § 8 гл. II). Заметим, что при интерпретации следует учитывать, что рассматриваемые элементы не являются независимыми: содержание труда, например, нельзя считать «очищенным» от влияния зарплаты, ибо в среднем более содержательная работа выше оплачивается и т.д. Следует также учитывать, что речь идет об *оценках* элементов, а связь между элементом и оценкой носит сложный, опосредствованный

характер. Например, нет прямой зависимости между удовлетворенностью зарплатой и ее величиной (в наших исследованиях было установлено наличие U-образной зависимости[25]). Зависимости опосредствуются потребностями, притязаниями. Так, удовлетворенность зарплатой зависит не столько от ее «абсолютной» величины, сколько от достижения «нормы», в качестве которой, как удалось установить, выступает средняя прогрессивная референтной группы (для работников промышленных предприятий ею оказалась их социально-профессиональная группа). Во всяком случае нами установлена тесная корреляция между удовлетворенностью зарплатой рабочих данной группы и числом работников, получающих зарплату не ниже среднепрогрессивной[26].

Пример 25. Коэффициент λ, определенный выше, описывает влияние *X* на *Y*. Мы обозначим его $\lambda_{y/x}$. Аналогично

[127]

можно ввести коэффициент $\lambda_{x/y} = \frac{E_x - E_{x/y}}{E_x}$, который описывает влияние *Y* на X. λ несимметричен: вообще говоря $\lambda_{y/x} \neq \lambda_{x/y}$ Если из содержательного анализа ясно, что X может влиять на Y и Y на X, то целесообразно вычислить оба коэффициента. Например, удовлетворенность работой (Y), влияет на удовлетворенность специальностью (X) и наоборот. Поэтому мы вычисляем оба коэффициента, используя их для сравнения указанных влияний. Так, в конкретном исследовании рабочих Ильичевского судоремонтного завода (1974г.) нами было получено такое двумерное распределение обсуждаемых признаков:

Таблица 31

Связь между удовлетворенностью работой и удовлетворенностью специальностью

X	*Y*			$N(x_i)$
	y_1	y_2	y_3	
x_1	1105	30	110	1245
x_2	313	55	62	430
x_3	35	4	36	75
$N(y_j)$	1453	89	208	1750

Для этой корреляционной таблицы, оказывается, $\lambda_{y/x}$ = 0,073, а $\lambda_{x/y}$ = 0,057. Таким образом, можно предположить, что удовлетворенность специальностью в большей мере влияет на удовлетворенность работой (предприятием), чем наоборот. Подчеркнем, что это утверждение относится к локальным условиям определенного, весьма специфического предприятия. Для изучения поставленного вопроса в целом необходимо провести дальнейшие исследования. В нашу задачу здесь входило ознакомление с идеей метода и техникой вычисления.

Упражнение 65. Вычислить $\lambda_{y/x}$ и $\lambda_{x/y}$ для таблицы из упражнения 62 самостоятельно.

Пример 26. *Энтропийный анализ социальных структур*.

В шестидесятые годы О.И. Шкаратан с группой сотрудников изучал социальную структуру современного промышленного предприятия. Результаты теоретического анализа, базирующегося на значительном эмпирическом материале, изложены в книге «Проблемы социальной структуры рабо-

[25] Аналогичный характер имеет зависимость между удовлетворенностью образованием и фактическим образованием (обследовались работники промышленных предприятий г. Одессы).

[26] *Максименко В. С., Попова И. М.* Заработная плата как фактор стимулирования трудовой деятельности.— В кн.: Проблемы экономики моря и мирового океана. Одесса, 1973, № 2

[128]

чего класса СССР» (М., 1970). Совместно с И.Н. Тагановым О.И. Шкаратан предпринимал попытки использования количественного метода для изучения указанной структуры. Одна из них, связанная с применением энтропийного анализа, была изложена в журнале «Вопросы философии» (1969, №5) и привлекла внимание социологов, интересующихся использованием количественных методов в социальных исследованиях. Рассмотрим ее суть применительно к фактически реализованной исследователями программе, но с использованием обозначений предыдущих параграфов.

Основная задача, которая решалась авторами с помощью энтропийного анализа, состояла в выделении свойств (признаков), определяющих неоднородность изучаемой социальной структуры. Задача рассматривалась в трехмерном пространстве, т.е. из гипотетического набора значимых признаков (он был составлен на основе предварительного анализа, сюда вошли такие характеристики, как образование, квалификация, пол, профессия и т.д. – всего 27 признаков) авторы выделяли каждый раз тройку признаков, набор которых давал различные пространства. Всего таких пространств можно выделить $C_{27}^{3} = 2925$.

Логика исследования такова. Каждый индивид данной совокупности является носителем различных признаков. Пусть он обладает *i*-м значением признака X, *j*-м – У, *r*-м – Z (в соответствии с ограничением, принятым авторами, мы рассматриваем пространство, определяемое признаками X, Y, Z). Информацию об одном индивиде можно рассматривать как вектор в данном пространстве. Совокупности из N рассматриваемых индивидов соответствует совокупность N векторов. Из всех возможных пространств (наборов признаков) нужно выделить такое, в котором векторы лежат наиболее плотными группами (набор признаков наиболее резко дифференцирует совокупность индивидов). Для отыскания таких пространств и был применен энтропийный анализ.

Неопределенность заполнения пространства векторами определяется величиной

$$E = -\sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{r=1}^{m} P_{ijr} \log P_{ijr} ,$$

где $P_{ijr} = \dfrac{N_{ijr}}{N}$ (здесь N_{ijr} – число индивидов, у которых X = x_i, Y = y_i, Z = z_i); $i = \overline{1,k}$; $j = \overline{1,l}$; $r = \overline{1,m}$

[129]

Если векторы равномерно заполняют пространство, то

$$N_{ijr} = \frac{N}{klm},\ P_{ijr} = \frac{1}{klm},\ E_{\max} = \log klm$$

Рассмотрим величину $\alpha = \dfrac{E_{\max} - E}{E_{\max}}$. Так как $0 \le E \le E_{\max}$, то $0 \le \alpha \le 1$, причем α = 0 соответствует E = E_{max}, т.е. отсутствию неоднородности в распределении векторов (отсутствию дифференциации общности), а α = 1 соответствует E = 0, т.е. максимальной неоднородности (максимальной дифференциации).

Очевидно, разным пространствам соответствуют различные α и формально задача сводится к отысканию пространства с максимальным α.

Упражнение 66. Показать, что

$$\alpha = \frac{N \log \dfrac{klm}{N} + \sum_{i}\sum_{j}\sum_{r} N_{ijr} \log N_{ijr}}{N \log klm}$$

На эмпирическом материале ленинградских социологов величина α оказалась максимальной для набора признаков «профессия – квалификация – образование». Именно в этом пространстве векторы лежат наиболее плотными группами, данный набор признаков наиболее резко дифференцирует изучаемую социальную общность.

8. Некоторые другие коэффициенты

В данном параграфе мы рассмотрим несколько статистических коэффициентов, которые не получили в социальных исследованиях такого широкого распространения, как, скажем, *r*, *ρ*, *T* и даже η). Однако в социологической литературе уже встречаются упоминания об их использовании отдельными авторами.

Мы считаем целесообразным рассмотреть определения, проанализировать их и привести примеры вычисления некоторых таких коэффициентов[27]. С одной стороны, это покажет читателю, что диапазон используемых методов значительно шире, чем может представиться по основной массе публикаций, с другой, позволит более свободно ориентироваться в научных статьях.

[130]

g – *коэффициент Гудмана (для номинальных шкал)*

Коэффициент Гудмана не является симметричным: $g_{yx} \neq g_{xy}$ Если мы рассматриваем X как независимый (факторный) признак, то его влияние на Y описывается с помощью коэффициента

$$g_{yx} = \frac{\sum_{i=1}^{k} \max N_{ij} - \max N(y_i)}{N - \max N(y_j)}, \text{ (II,8,1)}$$

где *max* $N(y_j)$ – максимальный маргинал зависимого признака, а *max* N_{ij} – максимальная частота в *i*-ой строке корреляционной таблицы.

Если данному X соответствует определенный Y, то в строке лишь одна частота с соответствующим маргиналом, искомая сумма максимумов обращается в N, следовательно, $g_{yx} = 1$

Если признаки независимы, то $N_{ij} = \frac{1}{N} N(x_i) N(y_j)$, как мы видели, и максимальная частота в *i*-ой строке там, где максимален Y-маргинал, т.е.

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

Теперь $g_{yx} = 0$. Итак, $0 \leq g_{yx} \leq 1$. Аналогично определяется g_{xy}, описывающий влияние Y на X.

Коэффициенты Гудмана целесообразно использовать, если из содержательных соображений ясно, что X может влиять на Y (и наоборот) и это влияние, вообще говоря, не симметрично.

В тех случаях, когда X не может влиять на Y (например, X – квалификация, Y – возраст), следует вычислять только g_{xy} (возраст влияет на квалификацию).

Упражнение 67. Рассчитать g_{yx} и g_{xy} для следующей корреляционной таблицы:

X	Y			$N(x_i)$
	y_1	y_2	y_3	
x_1	20	0	0	20

[27] Обзор ряда других коэффициентов можно найти в кн.: Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. М., 1977, гл. III, IV.

x_2	0	15	30	45
$N(y_j)$	20	15	30	65

Ответ: $g_{yx} = 0{,}57$; $g_{xy} = 1$.

[131]

Интерпретируем результат:

Задание Y однозначно определяет X (см. таблицу). Соответственно $g_{yx} = 1$; но задание X не определяет еще Y (например, если $X = x_2$, то Y может быть и y_2, и y_3), соответственно $g_{yx} < 1$

Упражнение 68. 1. Записать любую диагональную таблицу и убедиться, что $g_{yx} = g_{xy} = 1$.

2. Сконструировать таблицу, для которой $g_{xy} < 1$, а $g_{yx} = 1$, Интерпретировать результаты расчета по аналогии с предыдущим.

Заметим, что выполнение этих несложных упражнений помогает уяснить смысл и различие коэффициентов g_{yx} и g_{xy}. Отметим также предлагаемый Б. Миркиным подход к обработке социологической информации[28], который может быть использован даже для случая номинальных шкал. В качестве меры близости признаков рассматривается мера близости разбиений общности, осуществляемых этими признаками.

Коэффициент близости разбиений

Обобщим формулу для меры близости между двумя разбиениями на случай корреляционной таблицы $k \times l$. В качестве исходной возьмем формулу, приводимую Б.Г. Миркиным и Л.Б. Черным в статье «Об измерении меры близости между различными разбиениями конечного множества объектов»[29].

Если R и S два разбиения множества из N элементов и R разбивает его на *m*, а S на *n* классов, причем в *i*-ом классе $|R_i|$ элементов, а в *j*-ом $|S_i|$ элементов, то мера близости разбиений

$$d(R,S) = \frac{1}{2}\sum_i \left|R_i\right|^2 + \frac{1}{2}\sum_j \left|S_j\right|^2 - \sum_i \sum_j \left|R_i \cap S_j\right|^2$$

(Здесь $R \cap S$ – пересечение классов R и S).

Так как $d_{\max} = \frac{1}{2}N(N-1)$, нормированная мера $\delta = \frac{2d}{N(N-1)}$, причем $0 \le \delta \le 1$, где $\delta = 0$ соответствует

[132]

максимальной связи, $\delta = 1$ минимальной (отсутствие связи).

Для корреляционной таблицы $\{N_{ij}\}$: признак X осуществляет разбиение общности N на *k* классов x_i, в каждом из которых $N(x_i)$ элементов: признак Y на *l* классов y_i, в каждом из которых $N(y_j)$ элементов.

Так как $N_{ij} = N(x_i) \cap N(y_j)$, то мера близости двух рассматриваемых разбиений

$$\delta(x,y) = \frac{1}{N(N-1)}\left[\sum_{i=1}^{k} N^2(x_i) + \sum_{j=1}^{l} N^2(y_j) - 2\sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}^2\right] \quad \text{(II,8,2)}$$

Допустим, что мы исследуем некоторое разбиение, осуществляемое Y, и хотим выяснить значимость ряда признаков $X^{(p)}$ ($p = 1, 2, ...$) для выявления данного разбиения.

---

[28] *Миркин Б.Г.* Новый подход к обработке социологической информации. – В кн.: Измерение и моделирование в социологии. Новосибирск, 1969.

[29] Автоматика и телемеханика, 1970, №5.

Значимость $X^{(p)}$ будет тем большей, чем ближе разбиения, т.е. чем меньше $\delta(X^{(p)}, Y) \equiv \delta_p$. Таким образом, значимость признака $X^{(p)}$ по отношению к разбиению Y можно принять обратно пропорциональной расстоянию δ_p. Эту значимость («силу влияния») можно интерпретировать, следуя Б.Г. Миркину, как меру связи между признаками. Пусть, например, разбиение У – это социально-профессиональные группы, а $X^{(p)}$ – различные социально-демографические признаки (профессия, квалификация, образование, доход, место жительства и т.д.), вычисляя δ_p, мы можем определить значимость (влияние) различных $X^{(p)}$ для выявления Y-разбиения, выделить наиболее информативные признаки.

Рассматривалась и такая задача: пусть Y – это расселение работников по «зонам доступности предприятия»[30], а X(p) – некоторые социально-демографические признаки, значимые для расселения. Наиболее значимым признаком оказалась принадлежность к социально-профессиональной группе.

Рассмотрим еще раз вопрос о связи между удовлетворенностями заработной платой и работой, используя для ее характеристики обсуждаемую меру (см. пример № 14 § 1 этой главы).

[133]

Теперь $S = \dfrac{A = B - C}{D}$, где

$$A = \sum_{i=1}^{3} N^2(x_i) = 448^2 + 508^2 + 52^2 = 461472,$$

$$B = \sum_{j=1}^{3} N^2(y_j) = 682^2 + 97^2 + 229^2 = 526974,$$

$$C = 2\sum_i \sum_j N_{ij}^2 = 2(350^2 + 35^2 + 63^2 + 298^2 + 52^2 + 158^2 + 34^2 + 10^2 + 8^2) = 490972$$

$D = N(N-1)$=1015056, δ – 0,490.

Упражнение 69. Показать, что для таблицы 19 § 1 этой главы δ = 0,528.

В первом случае δ меньше, но так как связь пропорциональна 1/ δ то она больше, чем во втором; таким образом, сделанный ранее вывод (§ 1) подтверждается.

Δ-коэффициент (номинальные шкалы)

В работе И.А. Шкрабкиной и Г.И. Смирновой «Программа измерения тесноты связи между двумя признаками»[31] предлагается для измерения связи использовать модульный коэффициент Δ.

Наряду с корреляционной таблицей $\{N_{ij}\}$ рассмотрим таблицу $\{\tilde{n}_{ij}\}$, где $\tilde{n}_{ij} = \dfrac{N_{ij}}{N(x_i)}$ и введем величину $\bar{n}_j = \dfrac{1}{k}\sum_{p=1}^{k} \tilde{n}_{pj}$ Мерой связи, точнее, влияния X на Y, может служить

$$S = \sum_{i=1}^{k} \sum_{j=1}^{l} \left| \tilde{n}_{ij} - \bar{n}_j \right|$$

Покажем это. Если признаки независимы, то $\tilde{n}_{ij} = \dfrac{N(y_i)}{N}$, а $\bar{n}_j = \tilde{n}_{ij}$, т.е. рассматриваемая сумма обращается в нуль.

Логика измерения связи такова: если признаки незави-

---

[30] «Зона доступности предприятия» определяется временем, затрачиваемым работником на передвижение от места жительства до места работы. По нормам градостроительства выделяются четыре зоны: А (до 30 мин.), Б (от 30 до 45 мин.), В (от 45 мин. до часа), Г (свыше часа).

[31] Анализ социологической информации с применением ЭВМ, ч.1. М., 1973, с.143-157

[134]

симы, то при изменении X значение Y не должно меняться, т.е. числа индивидов $\widetilde{n}_{ij}$ с разными X при фиксированном Y должны быть примерно одинаковы, т.е. равными $\overline{n}_j$. Если же X влияет на Y, то $\widetilde{n}_{ij}$ должны отличаться от среднего $\overline{n}_j$.

Обсуждаемая сумма не является нормированной. В работе Шкрабкиной и Смирновой в качестве коэффициента при сумме предлагается использовать величину $\frac{1}{k}$. Однако как легко видеть, $\Delta' = \frac{S}{k}$ не является нормированной величиной.

Для нормировки необходимо найти максимальное значение суммы S. Оказывается, что оно достигается в случае полной связи (связь мы называем полной, если каждому X соответствует одно значение Y) и равно $\frac{2}{k}(m-1)(2k-m)$, где $m = \min(k,l)$ – меньшее из чисел k и l.

Таким образом, нормированный коэффициент, описывающий влияние X на Y:

$$\Delta_{yx} = \frac{k}{2(m-1)(2k-m)} \sum_{i=1}^{k} \sum_{j=1}^{l} \left| \widetilde{n}_{ij} - \overline{n}_j \right| \qquad \text{(II,8,3)}$$

Итак, $0 \le \Delta \le 1$, причем 0 соответствует отсутствию, а 1 – полной связи.

Аналогично

$$\Delta_{xy} = \frac{k}{2(m-1)(2k-m)} \sum_{i=1}^{k} \sum_{j=1}^{l} \left| \widetilde{\widetilde{n}}_{ij} - \overline{\overline{n}} \right|,$$

где

$\widetilde{\widetilde{n}}_{ij} = \frac{N_{ij}}{N(y_j)}$, а $\overline{\overline{n}}_i = \frac{1}{l} \sum_{p=1}^{l} \widetilde{\widetilde{n}}_{ip}$.

Все ранее рассмотренные здесь коэффициенты применимы даже для номинальных шкал. Перейдем к коэффициентам, которые используются при наличии упорядочения значений признаков.

γ-коэффициент Гудмана

По определению

$$\gamma = \frac{P-Q}{P+Q} \qquad \text{(II,8,4)}$$

[135]

где P – число пар объектов, у которых оба признака упорядочены в одинаковой последовательности, а Q – то же, но в обратной.

Пусть значения X и Y в корреляционной таблице выписаны в одинаковой последовательности. Величину P можно вычислить как сумму результатов умножения частот каждой

Таблица 32

Пример расчета γ-коэффициента Гудмана

X	Y			N(x_i)
	y_1	y_2	y_3	
x_1	35	15	5	55

x_2	5	25	15	42
$N(y_j)$	40	40	20	100

клетки на сумму частот, расположенных в клетках ниже и правее:

$$P = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}\left(\sum_{r=i+1}^{k}\sum_{s=j+1}^{l} N_{rs}\right) . \quad \text{(II,8,5)}$$

Это выражение, очевидно, совпадает с уменьшаемым в формуле (II,6,9). Q – сумма результатов умножения частот каждой клетки на сумму частот, расположенных ниже и левее ее:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{l} N_{ij}\left(\sum_{r=i+1}^{k}\sum_{s=1}^{j-1} N_{rs}\right) \quad \text{(II,8,6)}$$

(Q – вычитаемое в упоминавшейся формуле).

Если связь полная и прямая, то Q = 0 и γ = 1, если же полная и обратная, то P = 0 и γ = – 1. Итак, $-1 \le \gamma \le 1$

Положительный γ-коэффициент Гудмана показывает, насколько вероятно, что при увеличении значения одного признака увеличится значение другого (отрицательный – при увеличении одного – уменьшается значение другого).

Так как этот коэффициент в наших социологических исследованиях еще не получил распространения, приведем пример его вычисления для простейшей таблицы 32.

P = 35(25+15)+15·15+5(25+15)+25·15=1625

$\widetilde{Q}$ = 5(5+25)+15·5=225

γ = +0,76

[136]

Упражнение 70. Для таблицы 32 рассчитать γ -коэффициенты Гудмана. Ответ: 0,33; 0,44.

d-коэффициент Сомерса

По определению,

$$d_{yx} = \frac{P - Q}{P + Q + Y_0} \quad \text{(II,8,7)}$$

$$d_{xy} = \frac{P - Q}{P + Q + X_0} \quad \text{(II,8,8)}$$

где Y_0 – число пар объектов с одинаковыми значениями Y (но разными X), а X_0 – с одинаковыми X (но разными Y), P и Q определены выше, см. (II,8,5), (II,8,6).

Вообще говоря, $X_0 \neq Y_0$ (далее мы рассмотрим способ их вычисления), следовательно, коэффициент *d* не является симметричным: $d_{yx} \neq d_{xy}$ Его следует применять, когда из содержательных соображений ясно, что влияние X на Y и Y на X неодинаково.

В частности, *d* используется, если не имеет смысла влияние, скажем, X на Y (удовлетворенность работой X не может влиять на возраст Y, хотя, например, может влиять на квалификацию). При этом вычисляется, естественно, лишь один коэффициент: в рассмотренном примере – d_{xy}, описывающий влияние Y на X.

Перейдем к вычислению X_0, т.е. числа пар объектов с одинаковыми X (но разными Y).

Для вычисления X_0 найдем сперва вклад $_i$-ой строки (все объекты этой строки имеют одинаковые значения X, равные x_i):

$N_{i1}N_{i2} \ldots N_{ij} \ldots N_{il}$

Число пар с одинаковыми X, но разными Y в этой строке:

$$N_{i1}(N_{i2}+N_{i3}+\ldots+N_{il})+N_{i2}(N_{i3}+\ldots+N_{il})+\ldots+N_{il-1}N_{il}=\sum_{p=1}^{l-1}N_{ip}\sum_{q=p+1}^{l}N_{iq}$$

Вклад всех строк и составляет X_0:

$$X_0=\sum_{i=1}^{k}\sum_{p=1}^{l-1}N_{ip}\sum_{q=p+1}^{l}N_i$$

[137]

Аналогично:

$$Y_0=\sum_{j=1}^{l}\sum_{p=1}^{k-l}N_{pj}\sum_{q=p+1}^{k}N_{qj}$$

Замечание. Так как число пар с одинаковыми X и Y

$$Z_0=\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{l}N_{ij}(N_{ij}-1),$$

то

$$Z_0+Y_0+X_0+P+Q=\frac{N(N-1)}{2}$$

Это соотношение можно применять для контроля вычислений. Для таблицы 32 вычислим коэффициенты Сомерса:

$X_0 = 35\cdot(15 + 5) + 15\cdot5 + 5\ (25 + 15) + 25\cdot15 = 1350;$

$У_0 = 35\cdot5 + 15\cdot25 + 5\cdot15 = 625;$

$d_{yx} = +0{,}57$

$d_{xy} = + 0{,}53$

Близость полученных значений d_{xy} и d_{yx} можно интерпретировать как «симметрию» влияния X на Y и Y на X. Легко видеть, что $|d| \le 1$ во всех случаях, причем $d = 0$, если связи нет. Приведем один пример использования рассмотренных коэффициентов в прикладных исследованиях.

Коэффициент γ широко применялся эстонскими социологами Института истории АН ЭССР при изучении удовлетворенности трудовой деятельностью. Согласно данным Т. Китвеля, ранжировка по γ оценок различных элементов рабочей ситуации по степени их связи с удовлетворенностью работой на данном предприятии имеет следующий вид: 1) содержание труда (0,597); 2) заработная плата (0,365); 3) сплоченность коллектива (0,340).

Далее идут: отношения с администрацией, организация труда и т.д.[32]

Обратим внимание на то, что эта последовательность сходна с той, которая была получена ленинградскими («Человек и его работа») и немецкими[33] социологами, а также находится в согласии с нашими результатами.

В наших исследованиях использовались: коэффициент Чупрова *T*, вариационный размах оценок, энтропийная мера связи λ. Все три способа дали одну и ту же последова-

[138]

тельность элементов: содержание труда, организация труда, заработная плата, отношения с администрацией и т.д. Отметим, что указанную последовательность элементов мы получили как с помощью показателя двусторонней связи – коэффициента Чупрова, так и с помощью показателя односторонней (направленной) связи – энтропийной меры связи.

Использованный Китвелем коэффициент γ является мерой двусторонней связи. Представляется целесообразным также применение несимметричного коэффициента

[32] Китвель Т.О социально-психологических проблемах удовлетворенности трудом. Таллин, 1974, с. 75.

[33] Stollberg R. Arbeitszufriedenheit – theoretische und praktische probleme. Berlin, 1967, S.49

Сомерса, который, учитывает последовательность позиций на шкале удовлетворенности (в этом его несомненное преимущество перед и T, и λ) и является «направленным» (в отличие от T и γ). С его помощью можно описать влияние частных удовлетворенностей (т.е. различными элементами) на интегральную удовлетворенность работой.

Существуют также некоторые коэффициенты, которые разработаны для случаев, когда одна переменная измерена по номинальной, а вторая – порядковой или метрической шкале. Мы рассмотрим два из них.

Ранговый бисериалъный коэффициент[34]

Предназначен для случая, когда одна шкала номинальная дихотомическая, а вторая – порядковая. Его название связано с тем, что при этом есть как бы две серии данных: каждая серия для одного из значений дихотомической переменной.

Назовем ранговым бисериальным следующий коэффициент (формула пригодна при отсутствии объединенных рангов):

$$r_{\rho b} = \frac{2}{N}(\bar{y}_1 - \bar{y}_2) \quad \text{(II,8,9)}$$

где N – число объектов; $\bar{y}_1$ – средний ранг по признаку Y объектов, имеющих значение x_1 дихотомической переменной X; $\bar{y}_2$ – средний ранг объектов, имеющих значение x_2. Пусть, например, дана дихотомическая переменная X ($x_1 = 1$, $x_2 = 2$) и ранговая переменная Y:

признак X	1	2	1	2	1	1	2	2	1	1
признак Y	1	10	2	9	5	8	4	7	3	6

В первой строке стоят значения признака X, а во второй – ранги признака Y для некоторых 10 объектов. Выпишем ранги по Y для каждого значения признака X:

[139]

X	Y	
x_1=1	1, 2, 5, 8, 3, 6	$\bar{y}_1 = \frac{25}{6} = 4{,}167$
x_2=2	10, 9, 4, 7	$\bar{y}_2 = \frac{30}{4} = 7{,}500$

Точечно-бисериальный коэффициент корреляции[35]

Предназначен для изучения связи признаков, один из которых измерен в номинальной дихотомической, второй – в метрической шкале:

$$r_{rb} = \frac{\bar{y}_1 - \bar{y}^2}{\sigma_y N}\sqrt{\frac{(N-1)N(x_1)N(x_2)}{N}} \quad \text{(II,8,10)}$$

где $\bar{y}_1$ – среднее значение признака Y для объектов, имеющих значение x_1 а $\bar{y}_2$ – значение x_2 дихотомической переменной X; N (x_1) и N (x_2) – число объектов, имеющих значение x_1 и x_2 соответственно, N – число всех объектов, σ_y – среднее квадратическое отклонение для всех объектов. Аналогично предыдущему коэффициенту рассмотрим следующую таблицу:

Значения X	Значения Y
x_1=1	170; 140; 157; 152; 155; 160; 152
x_2=2	150; 160; 165; 183; 163; 168; 160; 157

[34] Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М., 1976, с. 165 – 167.

[35] Там же, с. 149-151. Отметим, что в таблице на с. 151 этой книги, видимо, опечатка в данных о росте, поэтому приведенные в ней результаты неверны.

$N(x_1)=7$, $N(x_2)=8$, $N=15$, $\bar{y}_1=155{,}14$, $\bar{y}_2=163{,}25$, $\sigma_y=9{,}31$, $r_{rb}=0{,}42$

Формула (II,7,10) представляет собой алгебраическое упрощение коэффициента *r* для случая, когда X – дихотомическая переменная, поэтому все расчеты можно было бы проводить и по формулам для *r*, например, (II,5,1) или (II,5,3). Обобщения этих коэффициентов (полисериальные коэффициенты) не получили широкого распространения.

[140]

Глава III

РЕГРЕССИИ

1. Основные понятия. Прямая регрессия. Криволинейные связи. Корреляционное отношение

Как отмечалось, при исследовании связи между двумя признаками находят распределение совокупности в виде корреляционной таблицы $\{N_{ij}\}$; тесноту связи характеризуют с помощью коэффициентов корреляции (глава II), а форму – с помощью уравнений регрессии, к рассмотрению которых мы и переходим.

Напомним, что каждому значению x_i, соответствует распределение *y:* y_j , N_{ij}, где $j=\overline{1,l}$. Такие распределения называют условными, условными называют и соответствующие средние

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

Полную среднюю $\bar{y}$ можно рассматривать как взвешенную сумму условных средних $\overline{y_i}$.

Упражнение 71. Показать, что $\bar{y}$, равное, по определению, $\frac{1}{N}\sum_{j=1}^{l} y_i N(y_j)$ равно $\frac{1}{N}\sum_{i=1}^{k}\overline{y_i}N(x_i)$.

Далее мы будем изучать связь $\overline{y_i}$, с x_i. Если ее можно представить в виде $\bar{y}=f(x_i)$, где $f(x)$ – некоторая известная функция, то уравнение $\overline{y_i}=f(x)$, следуя Гальтону, называют *уравнением регрессии У на X*, а соответствующую ему кривую – *кривой регрессии*[1]. *С* таким уравнением мы уже встречались в примере 42 (§1 главы II).

[141]

Аналогично (III,1,1) определяется условная средняя

$$\overline{x_i}=\frac{\sum_{i=1}^{i} x_i N_{ij}}{N(y_j)}, \qquad \text{(III,1,2)}$$

соответствующая y_j (III, 1,2).

[1] Индекс *x* показывает, что речь идет об условном среднем.