

СТАТИСТИЧЕСКОЕ СОГЛАСОВАНИЕ МЕР СВЯЗИ В АНАЛИЗЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОЙ ИНФОРМАЦИИ

Ростовцев П. С.

(Новосибирск)

Предлагается подход, позволяющий использовать нормирующее преобразование мер связи, основанное на единой интерпретации — статистической значимости показателей.

На первых этапах исследования сложных социально-экономических объектов при определении направлений более углубленного анализа часто используют коэффициенты связи между парами анализируемых показателей.

В качестве простейшего примера рассмотрим связь жилищных условий работников с их общими характеристиками*. Жилищные условия имеют признаки: x_1 — жил. площадь/чел. в семье; x_2 — число комнат/чел. в семье; x_3 — оценка жилищных условий (хорошие, средние, плохие). Общие характеристики работника: y_1 — пол; y_2 — возраст; y_3 — семейное положение. Получив матрицу коэффициентов связи, столбцы и строки которой соответствуют признакам этих множеств, по их величинам можно судить о перспективности дальнейшего изучения связи пар переменных. В данном случае признаки «пол» и «семейное положение» могут рассматриваться только в номинальных шкалах; «возраст», «жилоплощадь» и «число комнат» — как количественные или ранговые; оценка жил. условий — как ранговый или номинальный признак. Для таких разнотипных данных формирование сравнимых между собой коэффициентов представляет определенную трудность.

Для оценки связи признаков отдельно взятых типов существует большое количество коэффициентов. Можно измерить взаимосвязь номинального признака с количественным, рангового с ранговым и т. п. Таким образом, исследуя данные анкетных обследований, получаем коэффициенты взаимосвязи между признаками «пол» и «профессия», «профессия» и «зарплата», «зарплата» и «возраст» и т. д. Обзоры таких показателей имеются, например, в [1–6].

В ряде случаев для удобства анализа данных признаки разного типа рассматривают в низших шкалах либо подвергают процедурам оцифровки [7]. Наш подход предполагает использование разнообразных способов оценки связи; для их согласования применяется единый способ интерпретации коэффициентов — вычисление их статистической значимости. Это дает возможность понимать абстрактные значения показателей связи в едином ключе как вероятность случайно получить величину коэффициента, в условиях независимости большую, чем рассчитанная с помощью эмпирических данных. Поэтому согласование коэффициентов на такой основе обеспечивает одинаковое ощущение силы взаимосвязи независимо от типа данных. Аналогичен описанный в [6] метод для при-

* Для иллюстрации используются данные анкетного обследования работников Мореходного комплекса Сахалина, проведенного ИЭ ОПП и Институтом морской геологии ДВО АН СССР.

менения односторонних критериев значимости при построении коэффициентов связи на базе комбинаторного подхода.

Потребности исследователей при анализе социально-экономических данных не ограничиваются изучением связи простых переменных. Нередко необходимо рассматривать не только отдельные признаки, но и группы переменных. По анкетам можно изучать связь определенной группы психологических характеристик и показателей производительности и качества труда. В нашем примере как единое целое выступает пара «жилплощадь» и «число комнат». Изучение такой связи происходит методами регрессионного, многомерного дисперсионного анализа, канонической корреляции. Они позволяют обобщить указанный подход для таких сложных образований, как группа признаков.

Вычисление значимости. Пусть $m(x, y)$ — произвольный коэффициент, характеризующий взаимосвязь x и y ; t — случайное значение m в условиях независимости аргументов и случайности выборки. Значимость $a(x, y)$ определяется как вероятность получить t не меньше, чем $m(x, y)$

$$a(x, y) = P\{t \geq m(x, y)\}. \quad (1)$$

Для некоторых коэффициентов оба конца интервала их изменения соответствуют наибольшей зависимости. В частности, величина коэффициента корреляции, равная 1 и -1 , соответствует линейной взаимосвязи признаков. Для таких мер будем считать, что значимость определяется по двустороннему критерию

$$a(x, y) = 2\min\{P\{t \geq m(x, y)\}; P\{t \leq m(x, y)\}; 0,5\}. \quad (2)$$

Стандартные единицы измерения связи. Для исследователей, соприкасающихся с анализом статистических данных, привычными характеристиками оценок параметров распределения являются среднеквадратичное отклонение и стандартная ошибка. Известно, что в условиях нормального распределения признака x около 70% разброса x находится в пределах симметричного относительно среднего интервала $(\bar{x} - s, \bar{x} + s)$, около 95% — в пределах интервала $(\bar{x} - 1,96 s, \bar{x} + 1,96 s)$, аналогичное представление об оценках параметров моделей дают стандартные ошибки (здесь s — оценка среднеквадратичного отклонения x).

Преобразование величины значимости связи с помощью функции, обратной функции распределения нормального закона

$$z(x, y) = \Phi^{-1}(1 + a(x, y)), \quad (3)$$

позволяет перейти от меры связи m к величине, выраженной такими же соотношениями. Здесь преобразование Φ^{-1} носит нормирующий характер. Будем называть $z(x, y)$ стандартизацией коэффициента связи m .

Таким образом, $z(x, y)$ определяет отклонение от независимости в стандартных единицах. Нулевое значение z соответствует медиане t — в 50% случаев можно в условиях независимости получить t , большую чем m . Равенство $z=1$ показывает, что большая величина связи может быть случайно обнаружена в 16% ситуаций; вероятность $P\{z \geq 1,96\}$ составляет около 2,5%.

Стандартизация и аппроксимация. Для приближенного определения стандартных показателей полезно использовать известные формулы, связывающие различные виды распределения с нормальным. В ряде случаев это приводит к некоторым ограничениям на теоретическое распределение данных, как, например, двумерная нормальность при стандартизации коэффициента корреляции. Применение формул аппроксимации позволяет получать величины стандартизованных показателей для очень малых величин значимости, таких, при которых показатели приобретают эвристический характер, поскольку аппроксимация становится недоста-

точно точной. Тем не менее опыт использования автором этих коэффициентов свидетельствует о том, что они не утрачивают практического значения.

Рассмотрим методы стандартизации для некоторых типов данных. **Стандартизация на основе гипергеометрического распределения.** Речь пойдет о связи дихотомических признаков.

Пусть N объектов некоторого множества описываются признаками x и y , принимающими значения 0 и 1; N_x объектов имеют значение $x=1$; N_y — значение $y=1$; для K объектов одновременно $x=1$ и $y=1$; K может послужить основой для стандартизации ряда коэффициентов связи дихотомических признаков.

При независимых x и y и фиксированных N_x , N_y и объеме выборки N величина K имеет гипергеометрическое распределение со средним

$$E = N_x N_y / N \quad (4)$$

и дисперсией

$$S^2 = N_x (N - N_x) (N - N_y) N_y / N^2 (N - 1). \quad (5)$$

При этом распределение

$$\hat{f} = (K - E) / S \quad (6)$$

при больших N близко к нормальному с параметрами $(0, 1)$.

Значение $\hat{f}$, вычисленное на эмпирических данных, непосредственно может использоваться для измерения связи [6]. Эта величина представляет собой стандартизованную характеристику совпадения свойств. При полном совпадении и полном несовпадении x и y $\hat{f}$ принимает максимальное и минимальное значения, соответствующие функциональной связи признаков на данных, поэтому для наших целей следует взять

$$z(x, y) = \Phi^{-1}(2(\Phi(|\hat{f}|) - 1)). \quad (7)$$

Использование точного критерия. В случае малой выборки аппроксимация может быть существенно неточной и это приведет к большим искажениям стандартизованных коэффициентов.

Приведем пример такого искажения для дихотомических переменных.

Пусть $N_x = N_y = K = 1$ при объеме выборки $N = 10$. Значимость α , вычисленная с использованием точного критерия Фишера [1] согласно (2), оказалась равной $2/10$. Этому значению соответствует стандартизованная величина связи $z(x, y) = 0,84$; но z , рассчитанное на основе (6) и (7), близко к $2,78$. Такие различия на практике неприемлемы.

Переход к аппроксимационному критерию при использовании бинарных признаков следует осуществлять, когда выборка достаточно велика. По [8] эмпирическим критерием «большой» выборки является соотношение

$$S^2 > 9. \quad (8)$$

На малой выборке велика вероятность получить $\alpha = 1$. В этом случае стандартное значение оказывается равным $-\infty$. Его невозможно использовать при анализе данных на ЭВМ и целесообразно заменить на достаточно большое по абсолютной величине отрицательное число, которым ограничиваются значения аргумента в опубликованных таблицах нормального распределения, например $z = -5$. Точный критерий полезен также для анализа связи между признаками.

Стандартизация коэффициента корреляции. Известно, что нормальное распределение аппроксимируется выборочным коэффициентом корреляции R_{xy} при теоретически нормальном двумерном распределении x и y с помощью преобразования Фишера [8]

$$U = \frac{1}{2} \log((1 + R_{xy}) / (1 - R_{xy})). \quad (9)$$

Аппроксимация нормального распределения распределением χ^2

Показатель	Значимость				
	0,5000	0,1000	0,0100	0,0010	0,0005
1	—0,018	1,306	2,336	3,043	3,222
2	—0,012	1,295	2,324	3,047	3,232
5	—0,004	1,287	2,322	3,061	3,253
10	—0,001	1,285	2,322	3,071	3,266
20	—0,001	1,283	2,324	3,079	3,276
40	0,000	1,282	2,335	3,084	3,282
100	0,000	1,282	2,326	3,087	3,287
	0,000	1,281	2,330	3,09	3,29

В условиях независимости x и y это распределение будет иметь параметры $(0,1/(N-3))$. Таким образом, коэффициент корреляции стандартизуется по формуле

$$z(x, y) = \Phi^{-1}(2\Phi((N-3)|U|) - 1). \quad (10)$$

Как было отмечено, при использовании этого коэффициента мы не свободны от предположения о виде распределения генеральной совокупности. Поэтому во многих случаях полезно применять непараметрические критерии, например коэффициент ранговой корреляции Кендалла. Существует достаточно простое нормальное приближение этого коэффициента [3]. Стандартизация его аналогична стандартизации обычного коэффициента корреляции.

Стандартизация мер связи, основанных на критерии χ^2 . Ряд мер связи номинальных признаков функционально выражается через показатель [4]

$$X^2(x, y) = \sum_{i=1}^r \sum_{j=1}^c (N_{ij} - N_{i.}N_{.j}/N)^2 / (N_{i.}N_{.j}), \quad (11)$$

где c, r — число значений признаков x и y ; N_{ij} — число объектов, имеющих значение i признака x и j признака y ; $N_{i.} = \sum_{j=1}^c N_{ij}$; $N_{.j} = \sum_{i=1}^r N_{ij}$;

N — общее число объектов. Величина $X^2(x, y)$ характеризует отклонение таблицы сопряженности от таблицы, ожидаемой в случае независимости x и y . При больших N и независимых x и y величина X^2 имеет распределение χ^2 с $v = (c-1)(r-1)$ степенями свободы.

Распределение χ_v^2 при больших v аппроксимирует посредством преобразования

$$g = ((\chi_v^2/v)^{1/2} - (1 - 2/(9v)))/(2/(9v)) \quad (12)$$

нормальное распределение с параметрами $(0, 1)$ [2].

Таким образом, подстановка в указанную формулу величины X^2 и числа степеней свободы дает стандартизацию коэффициента связи.

При малых ожидаемых частотах величина критерия χ^2 часто оказывается сильно завышенной (не соответствующей распределению χ^2). Существующие методы «исправления» критерия позволяют частично устранить этот недостаток.

Масштабы искажения показателя z , обусловленные неточностью аппроксимации по формуле (12), показаны в табл. 1. Ее элементы получены подстановкой критических значений распределения χ^2 в указан-

ную формулу; последняя строка соответствует критическим значениям нормального распределения.

Для представленных в табл. 1 уровней значимости даже при $v=1$ ошибка составляет сотые доли стандартной единицы, что показывает применение формулы (12) и при малых v .

Связь между количественной и номинальной переменными. Пусть, как принято в одномерном дисперсионном анализе, совокупность объектов делится в соответствии со значениями номинального признака Y на k классов; X_{ij} — значение признака X на объекте j класса i ; $\bar{X}_i$ — среднее значение X на классе i ; $\bar{X}$ — на всей совокупности. Для оценки значимости связи в дисперсионном анализе применяется F -отношение

$$F = \left(\sum_i (\bar{X}_i - \bar{X})^2 N_i / (k - 1) \right) / \left(\sum_{i,j} (\bar{X}_i - X_{ij})^2 / (N - k) \right). \quad (13)$$

Этот критерий функционально увязан с такой мерой связи, как коэффициент детерминации.

В предположении независимости и теоретической нормальности совокупности на ЭВМ достаточно просто вычисляется значимость и, следовательно, стандартизованное значение этого коэффициента.

Для быстрого проведения расчетов можно учесть, что при больших N величина $(k-1)F$ в условиях гипотезы имеет распределение, близкое к распределению χ^2 , и, таким образом, стандартизованный показатель можно получить, пользуясь (12).

Связь номинального признака с системой количественных признаков. Если надо рассматривать систему количественных признаков как единое целое, естественно исследовать статистики, возникающие в многомерном дисперсионном анализе [1].

Пусть в отличие от случая одномерного дисперсионного анализа совокупность объектов описывается многомерными векторами $X = (X_1, \dots, X_m)$. Для оценки связи здесь нужно воспользоваться критерием Уилкса [8], представляющим собой обобщение однофакторного анализа для многомерных совокупностей. С этой целью оцениваются матрицы обобщенного внутригруппового разброса W и межгруппового разброса B . Критерий Уилкса при этом представляет собой отношение детерминант матриц

$$U = |W| / |W+B|. \quad (14)$$

Далее, аппроксимация Бартлетта

$$B(U) = -(N - (m+k+2)/2) \log(U) \quad (15)$$

позволяет воспользоваться для определения значимости распределением χ^2 с $m(k-1)$ степенями свободы. Распределение χ^2 , в свою очередь, легко стандартизуется.

Некоторые замечания. Использование одностороннего критерия. Список типов данных, связь между которыми достаточно просто представить в стандартных единицах, можно продолжить, используя для этого значимость статистик, возникающих в регрессионном анализе, анализе канонической корреляции. Все это не внесет принципиально нового в данную работу. На наш взгляд, приведенных примеров довольно, чтобы продемонстрировать, что практически любые показатели связи, для которых есть сведения об их теоретическом распределении, могут сравниваться на основе общей интерпретации.

Стандартизованные коэффициенты связи простых и сложных, однотипных и разнотипных данных для сопоставления удобно размещать в единой таблице. Подобные таблицы, конечно, грубы для анализа, однако дают возможность быстро сориентироваться в направлении поиска

Таблица 2

Взаимосвязь общих сведений о работниках с признаками жилищных условий

Показатель	Жилая площадь, x_1	Число комнат, x_2	Оценка жилищных условий, x_3
Пол, y_1	-0,04	0,11	2,01
Возраст, y_2	2,31	3,10	3,85
Семейное положение, y_3	10,15	7,23	0,09

существенных для исследования закономерностей. Целесообразно подготовить программное обеспечение ЭВМ для подобного анализа связи.

Все это верно для определенных выбранных моделей оценки связи. Переход к другой модели может дать иной результат.

Использование двустороннего критерия (2) лишает нас возможности понять направленность зависимости; при этом подходе измеряется лишь «сила» связи. Если изучаются только упорядоченные переменные, при стандартизации целесообразно применять односторонние критерии. Тогда положительные значения z будут соответствовать прямой, а отрицательные — обратной зависимости. Это прежде всего касается коэффициента корреляции, в том числе ранговой, мер взаимосвязи бинарных признаков. Последние позволяют для стандартизации использовать непосредственно формулы (8) и (9).

Примеры использования стандартизации. Вернемся к примеру задачи, сформулированной в начале статьи. Не вдаваясь в подробности, приведем таблицу стандартизованных показателей связи общих сведений о работниках с признаками жилищных условий (табл. 2).

Для получения коэффициентов связи $z(x_1, y_1)$, $z(x_2, y_2)$, $z(x_3, y_2)$, $z(x_1, y_3)$, $z(x_2, y_3)$ использован F -критерий Фишера; $z(x_1, y_2)$, $z(x_2, y_2)$ — коэффициент Кендэлла; $z(x_3, y_1)$, $z(x_3, y_3)$ — критерий χ^2 . Величины связи $z(x_1, y_3)$ и $z(x_2, y_3)$, вычисленные на основе аппроксимационных формул, настолько велики, что в большинстве статистических таблиц нет соответствующих значений функции распределения нормального закона. Скорее всего в этом случае аппроксимация неточна по сравнению с соответствующим уровнем значимости.

Рассматривая табл. 2, можно, например, заключить, что с точки зрения обнаружения зависимости наиболее перспективны коэффициенты связи семейного положения с количественными характеристиками жилищных условий, а связь пола с ними не обнаруживается. В то же время есть некоторая значимая связь пола с оценкой жилищных условий. Таким же образом можно интерпретировать и другие коэффициенты. Дальнейшее исследование может быть более прицельным. Имеет смысл, например, рассмотреть таблицы сопряженности или дисперсионного анализа и т. д., изучить внешние факторы, определяющие эти связи.

Небольшие значения стандартизованных показателей необязательно свидетельствуют об отсутствии связи или даже о малой зависимости данных, а возможно, о неподходящей модели, использованной для их вычисления. Большие по абсолютной величине отрицательные значения этих показателей говорят о принципиальном характере такого несоответствия или независимости. В частности, применение в наших расчетах коэффициента корреляции вместо коэффициента Кендэлла дает незначимые величины коэффициентов.

Многомерный дисперсионный анализ позволяет в стандартных единицах оценить величину связи жилищных условий x_1 и x_2 как единого целого и семейного положения $y_3 - z(\{x_1, x_2\}, y_3) = 9,8$.

Связь признака «Наличие коммунальных удобств» с признаком «Семейное положение»

Удобства	Состоит в браке	Не состоит в браке	Разведен
Водопровод	0,41	-0,52	0,00
Газ	-2,51	2,26	0,73
Канализация	-1,13	0,18	0,98
Центральное отопление	-0,53	-0,01	0,72
Горячее водоснабжение	-0,18	0,36	0,00
Телефон	0,41	-0,20	0,35
Мусоропровод	0,00	0,00	0,18
Лифт	0,08	-0,33	0,00

Продемонстрируем использование одностороннего критерия при анализе таблиц сопряженности.

Эти таблицы характеризуют связь между неколичественными признаками. Социологами применяется также обобщение таблиц сопряженности для многозначных (неальтернативных) признаков, соответствующих вопросам вида «Что Вам нравится в Вашей работе?» или «Какими видами удобств обеспечено Ваше жилье?» Элементы таблиц сопряженности можно рассматривать как коэффициенты связи между бинарными признаками, соответствующими значениям признаков. При их анализе преимущественно используют сравнение процентных распределений по строкам или столбцам таблиц. Существует большое число коэффициентов, применяемых как характеристики связи значений. В частности, ими являются слагаемые в (11), по которой вычисляется статистика χ^2 ; применяется также коэффициент, вычисляемый по (6) [9]. Эти показатели отражают отклонение частот от ожидаемых в условиях независимости. Выше приведена табл. 3, полученная с помощью этой формулы в сочетании с использованием точного критерия Фишера.

Из этой таблицы видно, что наиболее существенной закономерностью является малая доля живущих в обеспеченных газом квартирах и состоящих в браке и относительно большая обеспеченных этим видом удобств среди не состоящих в браке. Доля обеспеченных газом в целом по совокупности равна 3,8%, среди состоящих в браке — 2,7%, среди не состоящих в браке и разведенных соответственно 11,1 и 6,8%. Таким образом, сопоставимые различия маргинальных долей и долей в группах необязательно характеризуют одинаково значимые отклонения. Это подтверждает целесообразность применения предложенного подхода.

ЛИТЕРАТУРА

1. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. М.: Мир, 1982.
2. Айвазян С. А., Енюков И. С., Мешалкин А. Д. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.
3. Кендэлл М. Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973.
4. Миркин Б. Г. Анализ качественных признаков и структур. М.: Статистика, 1980.
5. Řehák J., Řeháková B. Základní charakteristiky proměnných s konečným počtem hodnot a distanční analýza jejich rozložení//Sociologický časopis. 1979. V. 15.
6. Lerman I. C. Combinatorial Analysis in the Statistical Treatment of Behavioural Data//Quality and Quantity. 1980. V. 14. № 3.
7. Интерпретация и анализ данных в социологических исследованиях. М.: Наука, 1987.
8. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М.: Наука, 1983.
9. Haberman Shelby J. Analysis of Qualitative Data. V. 1. Introductory Topics. L., 1978.

Поступила в редакцию
29 V 1990