

СРАВНИТЕЛЬНЫЙ АНАЛИЗ «ПЕРВЫХ» КРИТЕРИЕВ ВЫЯВЛЕНИЯ СВЯЗИ МЕЖДУ КАЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

Проведен сравнительный анализ использования «первых» критериев оценки связи между категориальными переменными. Показана согласованность этих критериев на примере выявления связи между типом авиарейса и числом крушений.

Ключевые слова: коэффициент ассоциации, коэффициент коллигации Юла, коэффициент контингенции, быстрый критерий, оценки связи.

Признак – это свойство, характерная черта явления, подлежащая статистическому изучению. Признаки в зависимости от своей природы классифицируются на количественные и качественные (атрибутивные). Признаки, отдельные значения которых различаются по величине, называются *количественными* (например, возраст, рост, вес). Качественные признаки выражают существенное неотъемлемое свойство предмета. Противоположные качественные признаки называются *альтернативными* (например, мужчина – женщина, ребенок – взрослый). Любой качественный признак можно свести к альтернативному (например, студент, обучающийся на «отлично», – студент, не обучающийся на «отлично», полет прошел в штатном режиме – возникла аварийная ситуация).

Как известно, основная задача статистики состоит в выявлении связи между признаками независимо от того, какую природу они имеют. Если связь имеется и она значимая, то вопрос о существовании связи влечет за собой вопрос о силе данной связи. Методов измерения тесноты взаимосвязи довольно много, но все они условно делятся на параметрические и непараметрические.

Непараметрические методы были первыми из методов измерения тесноты взаимосвязи количественных признаков. Эти методы применяются для измерения тесноты связи качественных и альтернативных признаков, а также количественных признаков, распределение которых отличается от нормального распределения. Непараметрические методы учитывают направление изменения значений признаков, но не зависят от того, насколько интенсивно колеблются значения результативного признака при изменении факторного признака, что позволяют сделать параметрические методы.

Для измерения связи альтернативных признаков применяются коэффициент ассоциации и коэффициент контингенции. Для расчета этих и других показателей применяется следующая матрица взаимного распределения частот (табл. 1).

Таблица 1

Таблица сопряженности 2×2

Категории	B_1	B_2	Всего
A_1	f_{11}	f_{12}	f_{10}
A_2	f_{21}	f_{22}	f_{20}
Всего	f_{01}	f_{02}	f_{00}

В табл. 1 приняты следующие обозначения: f_{11} – число элементов выборки, обладающих признаками A и B одновременно; f_{12} – число элементов выборки, обладающих признаком A , но не обладающих признаком B ; f_{21} – число элементов выборки, обладающих признаком B , но не обладающих признаком A ; f_{22} – число элементов выборки, не обладающих ни одним из признаков.

Таким образом, f_{ij} можно определить как наблюдаемую частоту респондентов или каких-то определенных признаков, попавших в ячейку (A_i, B_j) . Символ f_{i0} обозначает частную сумму всех значений признака, попавших в категорию A_i , аналогично символ f_{0j} обозначает сумму значений сопряженного ему признака, относящихся к категории B_j , а символ f_{00} – общий итог всех рассмотренных случаев. Представим все сказанное выше на языке формул [1]:

$$f_{i0} = \sum_j f_{ij}, \quad f_{0j} = \sum_i f_{ij}, \quad f_{00} = \sum_i f_{i0} = \sum_j f_{0j} = \sum_i \sum_j f_{ij}.$$

Суммарные частоты по строке или по столбцу называются *маргинальными частотами*. Их значения важны, так как они позволяют оценить распределение частот в отдельных столбцах и строках таблицы. Сумма маргинальных частот равна объему выборки $f_{00} = n$, а их распределение представляет собой одномерное распределение переменной, образующей строки или столбцы таблицы.

Критериев выявления связи между качественными признаками существует достаточно много. Среди них можно выделить критерий Фишера, который позволяет сделать максимально точный вывод о наличии или отсутствии связи, но расчет которого затруднен, классический критерий χ^2 и G -критерий Вульфа, точность которых можно увеличить, используя поправку Йетса. Однако суть этих критериев состоит в том, что все они не выявляют связь между переменными, а только проверяют заранее выдвинутую гипотезу о наличии связи. Поэтому сначала следует воспользоваться одним из так называемых «первых» критериев, которые позволяют предварительно оценить возможность существования связи между переменными и выдвинуть правдивую нулевую гипотезу. К таким критериям относятся коэффициент ассоциации, коэффициент коллигации Юла, коэффициент контингенции и быстрый критерий оценки связи.

Коэффициент ассоциации. Метод измерения связи с помощью четырехклеточной таблицы был предложен английскими статистиками Э. Дж. Юлом и М. Дж. Кендаллом. Числитель этого коэффициента, называемого коэффициентом ассоциации, представляет собой определитель исследуемой четырехклеточной таблицы сопряженности, а знаменатель – сумму тех же произведений, разность которых стоит в числителе:

$$Q = \frac{f_{11}f_{22} - f_{12}f_{21}}{f_{11}f_{22} + f_{12}f_{21}}. \quad (1)$$

Если признаки A и B независимы, то $Q = 0$. В случае полной связи между признаками $Q = \pm 1$.

Самым существенным недостатком использования коэффициента ассоциации является его непригодность в случае, если одна из частот равна 0. При этом коэффициент ассоциации всегда равен ± 1 , а это можно ложно трактовать как полную зависимость между признаками. Поэтому если хотя бы одна из частот в таблице сопряженности равна 0, то критерий ассоциации применять нельзя и вместо него используется коэффициент контингенции.

Коэффициент ассоциации в плане расчета является относительно простым показателем сопряженности величин. Он применяется к вариации двух качественных признаков, распределенных по двум группам. Его значения принадлежат промежутку от -1 до $+1$. Интерпретируется он следующим образом: чем ближе коэффициент по абсолютной величине к 1, тем теснее связь (прямая или обратная). Считается, что если значение коэффициента ассоциации достигает 0,3, то это свидетельствует о возможной существенной связи между признаками. Однако можно сказать, что этот показатель очень либерально оценивает тесноту связи, завышая ее.

Если предположить, что суммарная частота f_{00} достаточно велика, то величина Q имеет нормальное распределение с дисперсией [2]:

$$D(Q) = \frac{1}{4}(1 - Q^2) \left(\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}} \right), \quad (2)$$

благодаря чему можно найти приближенные доверительные границы для Q . Сравнение Q с полученным значением дисперсии (с учетом масштаба $\sqrt{D(Q)}$) позволяет получить хотя бы первое приближение по оценке связи. Но так как в формуле (2) присутствуют дроби, то, исходя из области определения, дисперсию в виде конечного числа можно вычислить только в том случае, если все значения в ячейках таблицы сопряженности отличны от нуля.

Коэффициент коллигации Юла. Этот способ выявления связи между переменными мало известен и практически не используется.

Существует выражение, позволяющее найти его приближенное значение, зная значение коэффициента ассоциации Q :

$$Q = \frac{2K_{Ю}}{1 + K_{Ю}},$$

где $K_{Ю}$ – коэффициент коллигации Юла.

Статистика коэффициента коллигации имеет вид

$$K_{Ю} = \frac{\sqrt{f_{11}f_{22}} - \sqrt{f_{12}f_{21}}}{\sqrt{f_{11}f_{22}} + \sqrt{f_{12}f_{21}}},$$

а дисперсия вычисляется по формуле

$$D(K_{Ю}) = \frac{1}{16}(1 - K_{Ю}^2) \left(\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}} \right).$$

Коэффициент контингенции (сходства). Этот коэффициент используется для выявления связи между признаками, не имеющими количественного выражения (альтернативная вариация, или изменчивость качественных признаков), и применяется только в том случае, когда исследуемые признаки имеют лишь две градации.

В математической формуле, описывающей статистику коэффициента контингенции, его числитель аналогичен коэффициенту ассоциации, т. е. представляет собой разность произведения частот на главной и побочной диагоналях, а знаменатель является квадратным корнем из произведения всех четырех маргинальных частот. В буквенных обозначениях табл. 1 имеем:

$$V = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{(f_{11} + f_{12})(f_{11} + f_{21})(f_{22} + f_{12})(f_{22} + f_{21})}}.$$

Связь считается подтвержденной, если коэффициенты $|Q| \geq 0,5$ и $|V| \geq 0,3$, в противном случае признаки A и B альтернативны. Это означает, что данный объект либо обладает признаком, либо не обладает (знак «минус» показывает наличие обратной связи при условии ее подтверждения). В случае $|Q| > 0,5$ и $|V| < 0,3$ ($|Q| < 0,5$ и $|V| > 0,3$) точного ответа о наличии или отсутствии связи нет.

Свойства коэффициента контингенции те же, что и у коэффициента корреляции. Коэффициент контингенции обращается в 0, если оба произведения в числителе точно уравниваются (что крайне маловероятно). Коэффициент контингенции равен $+1$ лишь в том случае, когда f_{12} и f_{21} одновременно равны 0 и -1 или когда f_{11} и f_{22} одновременно равны 0. В отличие от него коэффициент ассоциации равен 1 уже в случае, если одно из чисел в клетках таблицы оказалось равным 0, как это уже отмечалось ранее. Величина коэффициента контингенции всегда меньше величины коэффициента ассоциации.

На практике для проверки гипотезы о существовании взаимосвязи между изучаемыми признаками используется величина $\chi^2 = nV^2$, имеющая при отсутствии связи распределение χ^2 с $d = 1$ степенью свободы.

С учетом поправки на непрерывность статистика критерия контингенции имеет вид

$$\chi^2 = \frac{f_{00} \left(|f_{11}f_{22} - f_{12}f_{21}| - \frac{f_{00}}{2} \right)^2}{(f_{11} + f_{12})(f_{11} + f_{21})(f_{22} + f_{12})(f_{22} + f_{21})},$$

где f_{00} – общее число данных в таблице, $f_{00} = f_{11} + f_{12} + f_{21} + f_{22}$.

Если $\chi^2 > \chi_{\alpha}^2$, то зависимость признаков A и B с достоверностью α признается значимой.

Критерий $\chi^2 = nV^2$ применяется при $f_{00} \geq 40$ и $f_{11}, f_{12}, f_{21}, f_{22} \geq 5$ [3].

Быстрый критерий оценки связи. Предыдущие коэффициенты позволяют грубо оценить меру связи между качественными переменными и дают возможность получить хотя бы предварительное представление о наличии связи и ее величине, а это в свою очередь позволяет корректно сформулировать нулевую гипотезу о зависимости или независимости переменных и далее проверить ее с помощью более точных критериев. Но кроме грубых оценок связи, позволяющих предварительно оценить ее меру, существуют математически более тонкие и быстрые, хотя также носящие скорее вспомогательный характер, способы оценки связи. К таким оценкам относится быстрый критерий.

В зависимости от некоторых особенностей расположения и значений частот в таблице сопряженности применяются разные статистики.

При $f_{11} + f_{12} + f_{21} + f_{22} \geq 25$ и при условии $f_{11} + f_{12} = f_{21} + f_{22}$ или $f_{11} + f_{21} = f_{12} + f_{22}$ быстрым критерием оценки связи является величина

$$z = \frac{(f_{11} + f_{22}) - (f_{12} + f_{21})}{\sqrt{f_{11} + f_{12} + f_{21} + f_{22}}}. \quad (3)$$

Связь признаков в таблице сопряженности частот с достоверностью α признается значимой, если $z > u_{\alpha}$, где u_{α} – α -квантиль стандартного нормального распределения.

Если $f_{11} + f_{12} \geq 10$ при условии $f_{11} + f_{12} \ll f_{21} + f_{22}$ или $f_{11} + f_{21} \ll f_{12} + f_{22}$, то справедлив критерий

$$\tilde{z} = \frac{f_{11} - f_{12} + \frac{(f_{11} + f_{21} - f_{12} - f_{22})(f_{11} + f_{12})}{f_{11} + f_{12} + f_{21} + f_{22}}}{\sqrt{f_{11} + f_{12}}}. \quad (4)$$

Если $|\tilde{z}| > u_{\alpha}$, то связь признаков признается значимой [3].

По внешнему виду таблицы сопряженности частот можно увидеть следующую закономерность. Если связь между признаками является прямой, то частоты значений признаков сконцентрированы на главной диагонали, т. е. в ячейках f_{11} и f_{22} . При наличии обратной связи значения сосредоточены на побочной диагонали, т. е. в ячейках f_{12} и f_{21} . При отсутствии связи частоты практически равномерно распределены по всему полю таблицы.

Рассмотренные выше «первые» критерии выявления взаимосвязи имеют ряд преимуществ, к которым относятся простота, легкость в расчетах, получение быстрого ответа на вопрос, существует ли какая-то зависимость между исследуемыми переменными. Но эти преимущества влекут за собой и недостатки их использования, среди которых самыми явными являются неудовлетворительная точность полученных выводов, в основном – завышение оценки силы связи, и отсутствие точного и прямого ответа на вопрос, зависимы ли переменные, потому что «первые» крите-

рии не утверждают, а лишь предполагают отсутствие или наличие связи.

Пример. В последнее десятилетие в связи со все большей востребованностью и доступностью гражданского авиатранспорта появляется много чартерных рейсов. При этом существует точка зрения, что число авиакрушений чартерных рейсов больше, чем обычных – см. данные таблицы 2 [4]. Проверим с помощью всех вышеперечисленных способов выявления связи, действительно ли существует зависимость между типом рейса и количеством аварий.

Таблица 2

Таблица авиакрушений с жертвами за 1959–2007 гг. [4]

Тип рейса	Количество авиакрушений с жертвами		Всего
	1959–1997 гг.	1998–2007 гг.	
Обычный	341	74	415
Чартерный	39	4	43
Всего	380	78	458

Расчет коэффициента ассоциации. Так как в таблице сопряженности все ячейки с частотами отличны от нуля, то по формуле (1) можно вычислить значение коэффициента ассоциации:

$$Q = \frac{341 \cdot 4 - 39 \cdot 74}{341 \cdot 4 + 39 \cdot 74} = -0,358.$$

В данном случае знак «минус» говорит о том, что связь обратная, т. е. с течением времени количество авиакрушений уменьшается. Модуль значения коэффициента равен 0,358, т. е. больше 0,3, что говорит о возможной существенной связи между признаками «тип рейса» и «количество аварий».

Исходя из полученного значения коэффициента ассоциации определим дисперсию:

$$D(Q) = \frac{1}{4} \left(1 - (-0,358)^2 \right) \left(\frac{1}{341} + \frac{1}{74} + \frac{1}{39} + \frac{1}{4} \right) = 0,064,$$

$$\sqrt{D(Q)} = \sqrt{0,064} = 0,253.$$

Анализируя полученные данные, можно сделать промежуточный вывод о том, что связь между типом рейса и числом аварий скорее не существует, так как коэффициент $Q = -0,358$ недостаточно велик по сравнению со своими среднеквадратическим отклонением, т. е. не превышает его более чем в три раза.

Зная дисперсию, рассчитаем приближенный доверительный интервал для истинного значения связи, измеряемой с помощью коэффициента ассоциации Q : $Q \pm 1,96 \sqrt{D(Q)}$. В результате получим:

$$-0,358 \pm 1,96 \sqrt{0,064} \text{ или } (-0,854; 0,138).$$

Нетрудно заметить, что этот интервал накрывает значение 0, соответствующее независимости признаков. Следовательно, имеющиеся данные не дают оснований для отбрасывания возможности независимости. Таким образом, коэффициент ассоциации гово-

рит о возможной существенной связи между признаками, не исключая возможности их независимости.

Расчет коэффициента коллигации Юла. Коэффициент коллигации в данном случае составит

$$K_{Ю} = \frac{\sqrt{341 \cdot 4} - \sqrt{39 \cdot 74}}{\sqrt{341 \cdot 4} + \sqrt{39 \cdot 74}} = -0,185,$$

а дисперсия будет

$$D(K_{Ю}) = \frac{1}{16} (1 - (-0,185)^2) \left(\frac{1}{341} + \frac{1}{74} + \frac{1}{39} + \frac{1}{4} \right) = 0,018,$$

$$\sqrt{D(K_{Ю})} = 0,133.$$

По рассчитанным показателям можно сделать предварительный вывод о том, что между переменными «тип рейса» и «безопасность полета» связи не существует. Мы имеем право это утверждать, поскольку значение коэффициента коллигации невелико и больше своего среднеквадратического отклонения только в 1,39 раза.

Расчет коэффициента контингенции. По данным табл. 2 рассчитаем коэффициент контингенции:

$$V = \frac{341 \cdot 4 - 39 \cdot 74}{\sqrt{(341 + 74)(341 + 39)(4 + 74)(4 + 39)}} = \\ = \frac{-1522}{\sqrt{528925800}} = -0,066.$$

Абсолютное значение коэффициента коллигации $|V| = 0,066$ можно трактовать как отсутствие связи между переменными. Значение коэффициента ассоциации $|Q| = 0,358 < 0,5$ свидетельствует лишь о возможной существенной связи. Также подтверждается то, что величина коэффициента контингенции всегда меньше величины коэффициента ассоциации: $0,066 < 0,358$. Вывод об отсутствии связи очевиден. Но не стоит забывать о том, что коэффициент ассоциации и коэффициент контингенции являются предварительными критериями и делать по ним достоверные выводы нельзя. К тому же, как упоминалось выше, коэффициент ассоциации очень либерально оце-

нивает тесноту связи, завышая ее. Поправку на непрерывность при данном расчете также использовать нельзя, потому что, несмотря на то что объем выборки больше 40 ($f_{00} = 458$), частота в одной ячейке меньше 5. Поэтому необходимо воспользоваться одним из более точных критериев.

Быстрый критерий оценки связи. Первое условие, состоящее в том, что сумма всех частот должна быть не меньше 25, для данной задачи выполняется, а второе условие, что суммы частот по строкам или столбцам должны совпадать, не соблюдается. Поэтому формулу (3) применять нельзя.

Сумма частот первой строки $f_{11} + f_{12} = 415 > 10$, но она намного больше суммы частот второй строки: $415 \gg 43$, и сумма частот первого столбца также намного больше суммы частот второго столбца: $380 \gg 78$. Следовательно, формула (4) неприменима.

Таким образом, значение коэффициента ассоциации говорит лишь о возможной связи между переменными, коэффициенты коллигации и контингенции свидетельствуют об отсутствии связи, т. е. видна согласованность «первых» критериев.

Итак, практически все критерии показали независимость возникновения аварийных ситуаций и крушений самолетов от типа рейса при общей тенденции к снижению аварийности. Следовательно, при полете на чартерных и обычных рейсах риск потерпеть крушение одинаков и с течением времени он уменьшается.

Библиографические ссылки

1. Аптон Г. Анализ таблиц сопряженности. М. : Финансы и статистика, 1982.
2. Кендалл М. Дж., Стьюарт А. Статистические выводы и связи. М. : Наука, 1973.
3. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : Физматлит, 2006.
4. Статистика авиакатастроф, причины авиакатастроф [Электронный ресурс]. URL: <http://www.letaem-bez-straha.ru/otzyvy.html> (дата обращения: 21.10.2011).

А. Н. Antamoshkin, О. А. Bakaeva

COMPARATIVE ANALYSIS OF «THE FIRST» CRITERIA OF DIAGNOSIS OF RELATION AMONG QUALITATIVE VARIABLES

A comparative analysis of the use of the «first» criteria of estimation of relation among categorical variables is presented in the article. Compatibility of the criteria presented, on the example of revealing of relation of the type of a flight and the number of wreckages, is given.

Keywords: coefficient of association, coefficient of colligacy of Yule, coefficient of contingency, rapid criterion.

© Антамошкин А. Н., Бакаева О. А., 2011