УДК 519.235

О. А. Бакаева, В. Н. Щенников

Использование критерия χ^2 для выявления связи между качественными переменными на основе «идеальных» таблиц сопряженности

В статье рассмотрен классический вариант использования критерия Хи-квадрат, а также предложен оригинальный способ выявления связи между качественными переменными, в основе которого лежат «идеальные» таблицы сопряженности. Приведен пример, иллюстрирующий применение данного критерия.

Ключевые слова: критерий Хи-квадрат, таблицы сопряженности, частоты, идеальная зависимость, идеальная независимость.

O. A. Bakaeva, V. N. Shchennikov

Use of Criterion χ2 to Reveal Communication between Qualitative Variables on the Basis of "Ideal" Tables of Conjugation

In the article the classical variant of use of the Chi-square criterion is considered, and also is offered an original way to reveal communication between qualitative variables where "ideal" tables of conjugation are the basis. The example illustrating application of the given criterion is resulted.

Keywords: the Chi-square criterion, tables of conjugation, frequency, ideal dependence, ideal independence.

При работе с таблицами сопряженности самым эффективным способом обнаружения зависимости между переменными является критерий Хи-квадрат. Достаточно часто можно встретить его всевозможные модификации, самая популярная из которых – поправка Йейтса. Но данный критерий теряет свою «чувствительность» при работе с малыми значениями в ячейках таблиц сопряженности, поэтому на его точность полагаться уже не приходится. Однако сам метод сравнения наблюдаемых и ожидаемых частот, лежащий в основе χ^2-критерия, достаточно хорош, так как относится к числу непараметрических методов. Данные методы являются универсальными, потому все вычисления можно проводить независимо от вида закона распределения данных, и, следовательно, нет необходимости предварительно проверять случайную величину на нормальное или какое-либо другое распределение, что, несомненно, усложняет процесс выявления зависимости.

Однако за внешней простотой данного критерия скрываются довольно жесткие условия его использования. Перед тем, как рассчитывать значение χ^2, необходимо обратить внимание на некоторые особенности его применения для выявления взаимосвязи межу двумя категориальными переменными при работе с таблицами сопряженности 2 × 2.

Корректность проведения теста Хи-квадрат определяется тремя условиями:

1) случайный выбор наблюдений;

2) ожидаемые частоты $f_{ij} < 5$ должны встречаться не более чем в 20 % полей таблицы;

3) суммы по строкам и столбцам всегда должны быть больше нуля [2].

В основе χ^2-критерия лежит отклонение наблюдаемых частот, то есть фактических (O), от ожидаемых частот, которые являются результатом некоторого рода вычислений, то есть теоретических (E). Конкретно данное отклонение рассчитывается как сумма квадратов разности этих частот, выраженная в долях теоретической частоты. Это утверждение и является стандартной статистикой Хи-квадрат, имеющей следующий вид:

$$\chi^2 = \sum_{\text{по всем яячейкка}} \frac{(O - E)^2}{E} \qquad (1)$$

Две переменные считаются взаимно независимыми, если наблюдаемые частоты (f_o) в ячейках совпадают с ожидаемыми частотами (f_e), то есть значение χ^2 используется для оценки меры рассогласованности наблюдаемого и ожидаемого результата. Если, согласно нулевой гипотезе, ожидаемый результат будет сильно отличаться от наблюдаемых значений, значит стоит поставить под сомнение справедливость нулевой гипотезы H_0: «Между переменными нет зависимости». Если значение $\chi^2_{\text{расч}} < \chi^2_p$ (d), где p – вероятность, связанная с распределением χ^2, то нулевая гипотеза верна, иначе нулевая гипотеза отклоняется.

Но данный критерий, к сожалению, относится к критериям, имеющим лишь приближенное распределение χ^2. Поэтому его предпочтительно использовать, когда ожидаемые частоты в ячейках велики. Также он дает удовлетворительные выводы и при относительно малых ожидаемых частотах – порядка нескольких десятков. Но если частоты совсем малы, $f_{ij} \le 5$, то вероятность ошибки становится достаточно большой, что может привести к ложным выводам. Далее будет предложен способ, позволяющий получать достоверные выводы при работе с таблицами сопряженности с любыми частотами в ячейках. Он основан на отклонении наблюдаемых частот от «идеальных».

Определение 1. Таблица называется идеально зависимой, если значения на одной из диагоналей равны между собой и равны половине общих частот, остальные частоты равны 0, а маргинальные частоты равны $f_{00}/2$.

Определение 2. Таблица называется идеально зависимой 1-го рода, если значения на главной диагонали равны между собой и равны половине общих частот, то есть $f_{11} = f_{22} = f_{00}/2$, а остальные частоты равны нулю, то есть $f_{12} = f_{21} = 0$.

Таблица 1

Схема идеально зависимой таблицы сопряженности 1-го рода

	B_1	B_2	Всего
A_1	f_{11}	0	f_{11}
A_2	0	f_{22}	f_{22}
Всего	f_{11}	f_{22}	f_{00}

Определение 3. Таблица называется идеально зависимой 2-го рода, если значения на побочной диагонали равны между собой и равны половине общих частот, то есть $f_{12} = f_{21} = f_{00}/2$, а остальные частоты равны нулю, то есть $f_{11} = f_{22} = 0$.

Таблица 2

Схема идеально зависимой таблицы сопряженности 2-го рода

	B_1	B_2	Всего
A_1	0	f_{12}	f_{12}
A_2	f_{21}	0	f_{21}
Всего	f_{21}	f_{12}	f_{00}

Определение 4. Таблица называется идеально независимой, если значения в каждой ячейке равны между собой и равны $f_{00}/4$, а остальные частоты равны 0, тогда маргинальные частоты равны $f_{00}/2$.

Таблица 3

Схема идеально независимой таблицы сопряженности

	B_1	B_2	Всего
A_1	$f_{00}/4$	$f_{00}/4$	$f_{00}/2$
A_2	$f_{00}/4$	$f_{00}/4$	$f_{00}/2$
Всего	$f_{00}/2$	$f_{00}/2$	f_{00}

По внешнему виду этих таблиц отчетливо видно сильную связь между двумя переменными A и B. Из таблицы 1 следует, что все объекты, обладающие признаком A_1, обладают и признаком B_1. А вот объектов с признаком A_1 и обладающих признаком B_2 нет. Аналогично, нельзя встретить и объекты, одновременно обладающие свойствами A_2 и B_1. Исходя из таблицы 2, можно сделать вывод, что в данной выборке существует 2 группы объектов: одна обладает одновременно свойствами A_2 и B_1, а

другая – свойствами A_1 и B_2. Однако наблюдаемые на практике значения не столь однородны, поэтому данные таблицы можно считать «идеальными» в смысле зависимости признаков A и B.

Учитывая данные определения, для проверки нулевой гипотезы о независимости признаков A и B нужно объединить информацию о различиях между «идеальными» (И) и реальными наблюдаемыми (О) значениями. Если рассмотреть величину $(О-И)^2/И$, то получается новая статистика Хи-квадрат. Эту статистику целесообразно будет назвать по определению таблиц, данные которых она использует, «идеальной»:

$$\chi^2(\text{идеал}) = \sum_{\text{по всем ячейка}} \frac{(О - И)^2}{И} \qquad (2)$$

В данном случае распределение χ^2 (идеал) будет приблизительно таким же, что и распределение χ^2. Число степеней свободы для χ^2-распределения будет вычисляться аналогично критерию Хи-квадрат для таблиц 2 × 2 [1]:

d = (число ячеек) – (число ограничений, налагаемых на данные).

Если же цель задачи – выявление зависимости между двумя категориальными переменными, то число ограничений, налагаемых на данные, только одно. Заключается оно в том, что сумма ожидаемых значений должна быть равна сумме наблюдаемых значений.

Как и в случае с классическим χ^2-критерием, следует учитывать, что χ^2 является непрерывным распределением, а использовано оно для представления дискретного распределения. Поэтому в некоторых случаях необходимо использовать «поправку на непрерывность» – поправку Йейтса. Ее суть состоит в следующем: аппроксимация распределения статистики Хи-квадрат (идеал) может быть улучшена понижением абсолютного значения разностей между «идеальными» и наблюдаемыми частотами на величину 0,5 перед возведением в квадрат.

Тогда формула (2) будет выглядеть следующим образом:

$$\chi^2(\text{идеал}) = \sum_{\text{по всем ячейка}} \frac{\left(|О - И| - \frac{1}{2}\right)^2}{И} . \qquad (3)$$

Так как основная задача состоит в выявлении зависимости между признаками, то следует рассмотреть гипотезу H_1: реальные результаты должны быть близки к «идеальным» либо 1-ого рода, либо 2-ого рода, то есть подтверждать наличие связи между факторами, но могут ожидаться различия, обусловленные случайной изменчивостью. Тогда альтернативная гипотеза H_0: «Идеальные данные» не правдоподобны, а переменные абсолютно независимы. Так как идеальные данные подразделяются на 2 вида, то будет и два вида гипотез H_1:

H_1': реальные данные аналогичны идеально зависимым данным 1-ого рода, то есть признаки зависимы;

H_1'': реальные данные аналогичны идеально зависимым данным 2-ого рода, то есть тоже признаки зависимы.

Для получения результата следует проверить только одну из гипотез H_1' или H_1'', выбор которой определяется следующим образом. В таблице наблюдаемых данных нужно просуммировать частоты, стоящие на главной и побочной диагоналях. Если сумма частот на главной диагонали больше суммы частот на побочной диагонали, то есть $f_{11}+f_{22} > f_{12}+f_{21}$, то следует проверить гипотезу H_0'. Если сумма частот на побочной диагонали больше суммы частот на главной диагонали, то есть $f_{12}+f_{21} > f_{11}+f_{22}$, то целесообразнее проверить гипотезу H_0''. Если же суммы частот на главной и побочной диагоналях приближенно равны, то есть $f_{12}+f_{21} \approx f_{11}+f_{22}$, то следует проверить обе эти гипотезы. В качестве проверки можно использовать следующий принцип: наименьшая из статистик Хи-квадрат (идеал) должна соответствовать той идеальной таблице, с которой работают.

Для найденного значения статистики χ^2 (идеал) находим критическую точку распределения χ^2_p (d). Если χ^2 (идеал) $> \chi^2_p$ (d), то с вероятностью p признается зависимость между исследуемыми признаками. В противном случае связи нет.

Проиллюстрируем все вышесказанное на следующем примере.

Пример 1. По данным Управления Роспотребнадзора в Республике Мордовия в 2009–2010 гг. было зарегистрировано 75 случаев заболевания вирусом N1H1 («свиной грипп») среди взрослых людей (18–58 лет) и 45 случаев среди детей (до 18 лет). Среди взрослых было 9 летальных исходов, среди детей летальных исходов не было. Данные по заболеваемости и результатам лечения представлены в таблице 4.

Таблица 4

Таблица частот результатов лечения вируса N1H1 в Республике Мордовия в 2009–2010 гг.

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0–17 лет	0	45	45
Взрослые 18–58 лет	9	66	75
Всего	9	111	120

Очевидно, что пациентов с диагнозом «выздоровел» большинство – 111 из 120 человек. Необходимо знать, существует ли реальная зависимость между возрастом и результатом лечения, то есть выявить, существуют ли особенности лечения для какой-то отдельной возрастной категории. В данном примере классический критерий χ^2-квадрат применять нецелесообразно, так, частота в одной из ячеек равна 0, поэтому следует воспользоваться критерием Хи-квадрат для идеальных таблиц.

Сначала необходимо выяснить, какую из гипотез H_1' или H_1'' следует проверять. Для этого следует посчитать сумму частот на главной и побочной диагоналях: $f_{11}+f_{22} = 0 + 66 = 66$, $f_{12}+f_{21} = 45 + 9 = 54$. Сумма элементов на главной диагонали больше, чем на побочной, следовательно, необходимо проверить нулевую гипотезу по отношению к «идеальной» таблице 1-го рода. Но сначала необходимо построить идеальные таблицы 1-го и 2-го рода, а затем уже, если значения частот ≤ 5, рассмотреть те же самые таблицы с поправкой Йейтса, и в качестве контроля выбрать минимальное значение из χ^2 (идеал 1 рода) и χ^2 (идеал 2 рода) и сравнить его с критической точкой χ^2-распределения.

Таблица 5

Идеально зависимая таблица 1 рода: возраст-результат лечения

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0–17 лет	60	0	60
Взрослые 18–58 лет	0	60	60
Всего	60	60	120

Таблица 6

Идеально зависимая таблица 2 рода: возраст-результат лечения

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0–17 лет	0	60	60
Взрослые 18–58 лет	60	0	60
Всего	60	60	120

Таблица 7

Идеально независимая таблица: возраст-результат лечения

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0–17 лет	30	30	60
Взрослые 18–58 лет	30	30	60
Всего	60	60	120

Таблица 8

Идеально зависимая таблица 1 рода: возраст-результат лечения с поправкой Йейтса

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0–17 лет	59,5	0,5	60
Взрослые 18–58 лет	0,5	59,5	60
Всего	60	60	120

Таблица 9

Идеально зависимая таблица 2 рода: возраст-результат лечения с поправкой Йейтса

	Результат лечения		
Возраст	Летальный	Выздоровление	Всего
Дети 0–17 лет	0,5	59,5	60
Взрослые 18–58 лет	59,5	0,5	60
Всего	60	60	120

Маргинальные частоты, как видно, при использовании поправки Йейтса не меняются, в сумме частоты дают объем выборки $f_{00} = 120$.

По данным таблицы 8 найдем значение статистики χ^2 (идеал 1 рода):

$$\chi^2(\text{идеал 1 рода}) = \frac{(0-59{,}5)^2}{59{,}5} + \frac{(45-0{,}5)^2}{0{,}5} + \frac{(9-0{,}5)^2}{0{,}5} + \frac{(66-59{,}5)^2}{59{,}5} =$$

$$= 59{,}50 + 3960{,}50 + 144{,}5 + 0{,}71 = 4165{,}21.$$

По данным этой же таблицы найдем значение статистики χ^2 (идеал 2 рода):

$$\chi^2(\text{идеал 2 рода}) = \frac{(0-0{,}5)^2}{0{,}5} + \frac{(45-59{,}5)^2}{59{,}5} + \frac{(9-59{,}5)^2}{59{,}5} + \frac{(66-0{,}5)^2}{0{,}5} =$$

$$= 0{,}50 + 3{,}53 + 42{,}86 + 8580{,}50 = 8627{,}39.$$

Минимальное значение из χ^2 (идеал 1 рода) и χ^2 (идеал 2 рода) равно χ^2 (идеал 1 рода) = 4165,21. Учитывая, что $\chi^2_{0{,}05}(3) = 7{,}815$ несравнимо мало с χ^2 (идеал 1 рода), следовательно, уровень значимости $p < 0{,}05$, поэтому гипотезу о независимости возраста и результата лечения следует отвергнуть. Между данными факторами существует достаточно сильная зависимость.

Далее, исходя из последнего утверждения, выдвинем и проверим следующую нулевую гипотезу H_0: переменные «возраст» и «результат лечения» зависимы. По данным таблицы 8 найдем значение статистики χ^2 (идеал незав.):

$$\chi^2(\text{идеал незав}) = \frac{(0-30)^2}{30} + \frac{(45-30)^2}{30} + \frac{(9-30)^2}{30} + \frac{(66-30)^2}{30} =$$
$$= 30{,}00 + 7{,}50 + 14{,}70 + {,}50 = 43{,}20.$$

Так как χ^2 (идеал. незав.) = 43,20 > 7,815 = $\chi^2_{0,05}$ (3), то нулевую гипотезу о независимости признаков «возраст» и «результат лечения» опять следует отвергнуть. Таким образом, используя первый раз χ^2-критерий для идеальных таблиц сопряженности, мы получили ответ на вопрос, связаны ли данные. Использование повторно данного критерия, подтвердило полученный вывод о наличии связи между переменными.

Следовательно, так как между данными признаками имеется связь, то есть у людей различных возрастных категорий наблюдается различная эффективность лечения, то можно выявить категорию больных с меньшей эффективностью и в дальнейшем искать способы и методы, увеличивающие эффективность лечения и долю выздоровевших.

Использование критерия Хи-квадрат не требует больших расчетов, он прост для понимания и интерпретации полученных выводов, этим он удобен для людей, не связанных профессионально с математикой или статистикой. Но требования к его применению ограничивают возможности его использования. Построение идеальных таблиц сопряженности и сравнение их значений с реальными частотами позволяет расширить рамки применения данного критерия Хи-квадрат и сделать выводы о наличии связи между переменными в различных областях знаний.

Библиографический список

1. Аптон, Г. Анализ таблиц сопряженности [Текст] / Г. Аптон ; пер. с англ. Ю. П. Адлера. – М. : Финансы и статистика, 1982. – 143 с.

2. Кобзарь, А. И. Прикладная математическая статистика [Текст] / А. И. Кобзарь. – М. : Физматлит, 2006. – 816 с.

3. Крамер, Г. Математические методы статистики [Текст] / Г. Крамер. – М. : Мир, 1976. – 648 с.

4. Новиков, Д. А. Статистические методы в медико-биологическом эксперименте (типовые случаи) [Текст] / Д. А. Новиков, В. В. Новочадов. – Волгоград : Изд-во ВогГМУ, 2005. – 84 с.