УДК 001.891.222:36

ПРИМЕНЕНИЕ КРИТЕРИЯ ХИ-КВАДРАТ ДЛЯ АНАЛИЗА ЗАВИСИМОСТИ МЕЖДУ КАТЕГОРИАЛЬНЫМИ ПЕРЕМЕННЫМИ

С. В. Маленков[1)], Б. А. Бадак [2)]

[1)]Белорусский национальный технический университет, Беларусь, Минск, malenkovstas8@gmail.com
[2)]Белорусский национальный технический университет, Беларусь, Минск, badak.bazhena@bk.ru

В работе рассмотрены теоретико-эмпирические основы статистического критерия хи-квадрат и его применение на примере решения практико-ориентированной задачи, включающей анализ данных о зависимости между полом и выбором транспорта.

Ключевые слова: критерий хи-квадрат; практическая реализация; язык программирования C++.

APPLYING THE CHI-SQUARE CRITERION TO ANALYZE THE RELATIONSHIP BETWEEN CATEGORICAL VARIABLES

S. V. Malenkov[1)], B. A. Badak [2)]

[1)]Belarussian national technical university, Belarus, Minsk, malenkovstas8@gmail.com
[2)]Belarussian national technical university, Belarus, Minsk, badak.bazhena@bk.ru

The paper considers the theoretical and empirical foundations of the chi-square statistical criterion and its application using the example of solving a practice-oriented problem involving the analysis of data on the relationship between gender and choice of transport.

Keywords: chi-square criterion; practical implementation; C++ programming language.

Введение

Статистический анализ играет ключевую роль в проверке гипотез, особенно когда речь идёт о зависимости между категориальными переменными. Одним из наиболее распространённых методов анализа является критерий хи-квадрат, который позволяет оценить, существует ли статистически значимая взаимосвязь между переменными.

Критерий хи-квадрат (χ^2) – это статистический метод, предназначенный для проверки гипотезы о независимости двух категориальных переменных [1]. Данный метод позволяет определить, существует ли статистически значимая взаимосвязь между переменными, сравнивая наблюдаемые частоты с ожидаемыми значениями, рассчитанными при условии независимости переменных.

Теоретические и практические сведения

Метод основан на построении таблицы сопряжённости, где строки и столбцы представляют категории двух переменных. Каждая ячейка таблицы содержит наблюдаемую частоту – число случаев, принадлежащих соответствующей категории. Основное предположение критерия хи-квадрат заключается в том, что, если переменные независимы, то распределение частот в таблице будет соответствовать ожидаемым значениям, рассчитанным на основе общей структуры данных. Значение критерия хи-квадрат рассчитывается по следующей формуле: $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, где X^2 – значение критерия хи-квадрат, которое измеряет степень расхождения между наблюдаемыми и ожидаемыми частотами, $\sum_{i=1}^{r}$ $\sum_{j=1}^{c}$ - суммирование по всем строкам (i) и столбцам (j) таблицы сопряжённости, где r – количество строк, c – количество столбцов. O_{ij} – наблюдаемая частота в i-строке и j-столбце. Это фактическое значение, измеренное в ходе эксперимента или исследования. E_{ij} – ожидаемая частота для i-строки и j-столбца, которая вычисляется по формуле: $E_{ij} = \frac{rowTotal[i]+colTotal[i]}{total}$. В этой формуле row(col)Total[i] – итоговая сумма наблюдаемых частот в i-строке и j-столбце соответственно, a total – это общая сумма всех наблюдений в таблице сопряжённостью, а $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ – вклад каждой ячейки таблицы в общее значение хи-квадрат.

Формула используется для оценки гипотезы о независимости двух категориальных переменных. Если хи-квадрат превышает критическое значение, нулевая гипотеза о независимости переменных отвергается, что свидетельствует о наличии статистически значимой связи.

Число степеней свободы (df) при использовании критерия хи-квадрат определяется размером таблицы сопряжённости и вычисляется по формуле: $df = (r-1)\times(c-1)$, где r – количество строк таблицы, c – количество столбцов таблицы

Степени свободы показывают, сколько значений в таблице остаются независимыми после определения итоговых сумм по строкам и столбцам.

Для оценки результатов критерия хи-квадрат используется критическое значение, которое зависит от: уровня значимости (α) – вероятности допущения ошибки первого рода (обычно принимается равным 0.05 или 5%); числа степеней свободы.

Критическое значение определяет границу, выше которой можно утверждать, что различия между наблюдаемыми и ожидаемыми значениями статистически значимы. Эти значения находятся в таблице критических значений хи-квадрат или рассчитываются с использованием статистического программного обеспечения.

Гипотеза о независимости переменных (нулевая гипотеза) проверяется следующим образом:

если рассчитанное значение хи-квадрат превышает критическое значение, нулевая гипотеза отвергается, что свидетельствует о наличии статистически значимой зависимости между переменными.

Если рассчитанное значение хи-квадрат меньше или равно критическому значению, нулевая гипотеза не отвергается, и делается вывод об отсутствии зависимости между переменными.

Критерий хи-квадрат применим только к категориальным данным.

Ожидаемые частоты в каждой ячейке таблицы должны быть достаточно большими (обычно не менее 5). При нарушении этого условия результаты теста могут быть некорректными.

Метод не определяет силу и направление связи между переменными; он только проверяет наличие зависимости.

Критерий хи-квадрат широко используется в различных областях, включая социологию, медицину, маркетинг и другие дисциплины, где необходимо проверить гипотезы о независимости категориальных переменных.

Процесс применения критерия хи-квадрат описывается следующим алгоритмам:

1. Сбор данных и построение таблицы сопряжённости, где строки и столбцы представляют категории двух переменных.

2. Расчёт итоговых частот по строкам и столбцам, а также общей суммы наблюдений

3. Вычисление ожидаемых частот для каждой ячейки таблицы.

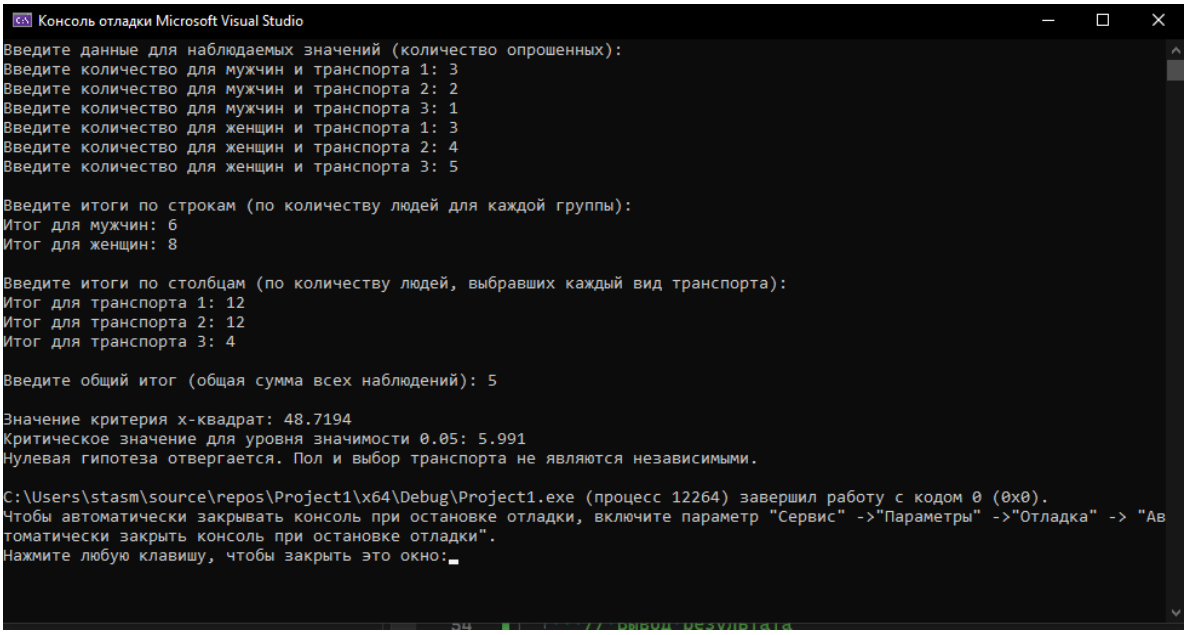
4. Расчёт хи-квадрат по приведённой выше формуле.

5. Сравнение полученного значения χ^2 с критическим значением, определяемым для заданного уровня значимости.

Если значение хи-квадрат превышает критическое, то гипотеза о независимости отвергается, а в противном принимается.

Для демонстрации применения критерия хи-квадрат рассмотрим задачу анализа зависимости между полом респондентов и выбором вида

транспорта. Данные представлены в виде таблицы сопряжённости, где строки – это категории пола (мужчины и женщины), а столбцы – виды транспорта. Наблюдаемые данные, а также итоги по строкам и столбцам используются для расчёта хи². Программа на языке C++ реализует этот процесс. Пример работы программы представлен на рисунке.



Пример работы программы

Заключение

Критерий хи-квадрат является мощным инструментом для анализа зависимости между категориальными переменными. Он широко применяется в социологических исследованиях, маркетинговых опросах, медицине и других областях. Рассмотренный пример демонстрирует, как с помощью простых вычислений и программного кода можно проверить гипотезу о независимости двух переменных, интерпретировать результаты и сделать выводы о характере их взаимодействия.

Библиографические ссылки

1. Теория вероятностей, математическая статистика и анализ данных: Основы теории и практика на компьютере. STATISTICA. EXCEL. Более 150 примеров решения задач: Учебное пособие. Москва: ЛЕНАНД. 2022.