КРИТЕРИЙ ХИ-КВАДРАТ ПИРСОНА

И. А. Марченко, К. А. Фесько,
студенты факультета международных экономических отношений и менеджмента
Учреждение образования Федерации профсоюзов Беларуси
«Международный университет «МИТСО», г. Минск
Научный руководитель:
Л. П. Фалько,
кандидат педагогических наук, доцент
доцент кафедры высшей математики
Учреждение образования Федерации профсоюзов Беларуси
«Международный университет «МИТСО», г. Минск

Результаты теории вероятностей и математической статистки широко используются в науке и практике анализа социально-экономических явлений и процессов. Теория вероятностей позволяет по одним вероятностям рассчитать вероятности, возникшие в результате исследования.

Основным моментом экономико-математического исследования является отображение реального процесса или явления в виде вероятностной математической модели. В процессе исследования используются два вида понятий:

1. Понятия, относящиеся к теории, – это вероятностная модель, а также математическое ожидание теоретического ряда.
2. Понятия, относящиеся к практике, – это выборочное наблюдение и выборочное среднее арифметическое значение.

С помощью вероятностной модели свойства, установленные по результатам анализа конкретной выборки, переносятся на генеральную совокупность. Чтобы перенести выводы о выборке на генеральную совокупность, используются предположения (гипотезы) о связи выборочных характеристик с теоретическими характеристиками генеральной совокупности.

Из вышесказанного следует актуальность темы исследования «Распределение "хи-квадрат"».

Цель данного исследования состоит в том, чтобы проанализировать применение распределения «хи-квадрат» на практике.

Для достижения цели сформулированы следующие задачи:

1. Изучить теоретические понятия распределения «хи-квадрат».
2. Проанализировать применение распределения «хи-квадрат» в задачах статического анализа данных.

Существует такое понятие, как критерий согласия – это статистический критерий проверки гипотезы о предполагаемом законе неизвестного распределения. Так как все предположения о характере того или иного распределения – это гипотезы, то они должны быть подвергнуты статистической проверке с помощью критериев согласия, которые дают возможность установить, когда расхождения между теоретическими и эмпирическими частотами следует признать несущественными, т. е. случайными, а когда – существенными (неслучайными).

Известны различные критерии согласия: Пирсона, Фишера, Смирнова и другие.

Критерий согласия Пирсона – наиболее часто употребляемый критерий для проверки простой гипотезы о законе распределения.

Для проверки гипотезы H_0 поступают так: разбивают область значений случайной величины X на m интервалов Δi и подсчитывают вероятности P_i попадания X в Δi по формуле $P\ (\alpha \leq X \leq \beta)\ =\ F$.

Существует еще один способ задания функции распределения «хи-квадрат» [1].

Рассмотрим случайную величину Y, распределенную по нормальному закону с параметром $M(Y) = a$ и средним квадратичным отклонением σ. То есть $Y \to N(a, \sigma)$.

Тогда, случайная величина $U = \frac{Y-a}{\sigma}$ называется стандартизированной случайной величиной, распределенной по нормальному закону с параметрами M (U) = 0, $\sigma_U = 1$, т. е. $U \to N(0, 1)$.

Квадрат стандартизированной случайной величины $U^2 = (\frac{Y-a}{\sigma})^2 = X^2$ называется случайной величиной X^2 одной степенью свободы.

Рассмотрим n независимых случайных величин $Y1, Y2, \ldots, Yn$, распределенных по нормальному закону с параметрами: математическими ожиданиями $a1, a2, \ldots, an$, и средними квадратическими отклонениями $\sigma1, \sigma2, \ldots, \sigma n$.

Образуем для каждой из них стандартизированную случайную величину:

$$U1 = \frac{Y_i - a_i}{\sigma_i}, i = \overline{1,4}$$

Сумма квадратов стандартизированных переменных:

$$X^2 = U_1^2 + U_2^2 + \cdots + U_n^2 = (\frac{Y_1 - a_1}{\sigma_1})^2 + (\frac{Y_2 - a_2}{\sigma_2})^2 + \cdots + (\frac{Y_n - a_n}{\sigma_n})^2$$

Называется случайно величиной X^2 с $V = n$ степенями свободы.

В статических таблицах число степеней свободы принято обозначать буквой V.

Плотность распределения случайно величины X^2имеет вид:

$$f(x^2) = \begin{cases} \frac{1}{2^{\frac{V}{2}} * \Gamma(\frac{V}{2})} * (x^2)^{\frac{V}{2}-1} * e^{-\frac{x^2}{2}}, \text{если } x^2 \geq 0 \\ 0, x^2 < 0 \end{cases},$$

где Γ – гамма-функция, или интеграл Эйлера 2-го рода вида:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} * e^{-x} dx$$

Гамма-функция является интегралом, зависящим от параметра α.

А распределение X^2зависит от одного параметра V – числа степеней свободы.

Функция распределения X^2 имеет вид:

$$F(x^2) = P(x^2 < x_0^2) = \begin{cases} \frac{1}{2^{\frac{V}{2}} * \Gamma(\frac{Y}{2})} * \int_0^{x^2} (x^2)^{\frac{V}{2}-1} * e^{-\frac{x^2}{2}} d(x^2), \text{если } x^2 \geq 0 \\ 0, \text{если } x^2 < 0 \end{cases}$$

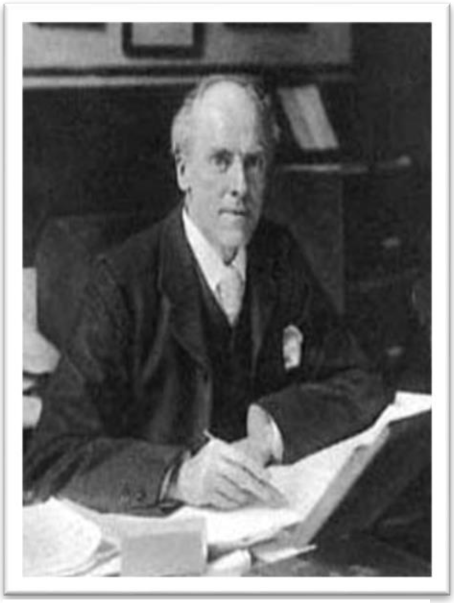

Рисунок 1 – Карл Пирсон

Критерий «хи-квадрат» для анализа таблиц сопряженности был разработан и предложен в 1900 году английским математиком, статистиком, биологом и философом, основателем математической статистики и одним из основоположников биометрики Карлом Пирсоном (1857 – 1936), показанном на рис. 1 [2].

Критерий χ^2 Пирсона – непараметрический метод, который позволяет оценить значимость различий между фактическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Выражаясь проще, метод позволяет оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей).

Критерий χ^2 применяется в двух целях:

1. Для сопоставления эмпирического распределения признака с теоретическим – равномерным, нормальным или каким-то иным.
2. Для сопоставления двух, трех или более эмпирических распределений одного и того же признака.

Критерий хи-квадрат может применяться при анализе таблиц сопряженности, содержащих сведения о частоте исходов в зависимости от наличия фактора риска. Четырехпольная таблица сопряженности показана в табл. 1.

Таблица 1 – Четырехпольная таблица сопряженности

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	A	B	A + B
Фактор риска отсутствует (0)	C	D	C + D
Всего	A + C	B + D	A + B + C + D

Таблицей сопряженности называется средство представления совместного распределения двух переменных, предназначенное для исследования связи между ними. Таблица сопряженности является наиболее универсальным средством изучения статистических связей, так как в ней могут быть представлены переменные с любым уровнем измерения. Такие таблицы получили наибольшее распространение при изучении социальных явлений и процессов: общественного мнения, уровня и труда жизни, общественно-политического строя и так далее.

Рассмотрим, как рассчитывается критерий χ^2 *на примере задачи:*

«Проводится исследование влияния курения на риск развития артериальной гипертонии. Для этого были отобраны две группы исследуемых: в первую вошли 70 человек, ежедневно выкуривающих не менее 1 пачки сигарет, во вторую – 80 некурящих такого же возраста. В первой группе у 40 человек отмечалось повышенное артериальное давление. Во второй артериальная гипертония наблюдалась у 32 человек. Соответственно, нормальное артериальное давление в группе курильщиков было у 30 человек (70 – 40 = 30) а в группе некурящих – у 48 (80 – 32 = 48)».

Заполняем исходными данными четырехпольную таблицу сопряженности (табл. 2).

Таблица 2 – Заполненная четырехпольная таблица сопряженности

	Артериальная гипертония есть (1)	Артериальной гипертонии нет (0)	Всего
Курящие (1)	40	30	70
Некурящие (0)	32	48	80
Всего	72	78	150

В полученной таблице сопряженности каждая строчка соответствует определенной группе исследуемых. Столбцы показывают число лиц с артериальной гипертонией или с нормальным артериальным давлением.

Задача, которая ставится перед исследователем: имеются ли статистически значимые различия между частотой лиц с артериальным давлением среди курящих и некурящих? Ответить на этот вопрос можно, рассчитав критерий хи-квадрат Пирсона и сравнив получившееся значение с критическим.

Для начала рассчитываем ожидаемые значения для каждой ячейки путем перемножения сумм рядов и столбцов с последующим делением полученного произведения на общее число наблюдений. Общий вид таблицы ожидаемых значений представлен в табл. 3.

Таблица 3 – Общий вид таблицы ожидаемых значений

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть(1)	(A + B) × (A + C) / (A + B + C + D)	(A + B) × (B + D) / (A + B + C + D)	A + B
Фактор риска отсутствует(0)	(C + D) × (A + C) / (A + B + C + D)	(C + D) × (B + D) / (A + B + C + D)	C + D
Всего	A + C	B + D	A + B + C + D

Подставляя исходные данные, получаем (табл. 4):

Таблица 4 – Ожидаемые значения с исходными данными

	Артериальная гипертония есть (1)	Артериальной гипертонии нет (0)	Всего
Курящие (1)	(70 × 72) / 150 = 33,6	(70 × 78) / 150 = 36,4	70
Некурящие (0)	(80 × 72) / 150 = 38,4	(80 × 78) / 150 = 41,6	80
Всего	72	78	150

На втором этапе находим критерий χ^2 *Пирсона по формуле:*

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Формула 1: критерий χ^2 *Пирсона*

где i – номер строки (от 1 до r);

j – номер столбца (от 1 до c);

O_{ij} – фактическое количество наблюдений в ячейке ij;

E_{ij} – ожидаемое число наблюдений в ячейке ij.

Подставляя исходные данные, получаем:

$$\chi^2 = \frac{(40 - 33{,}6)^2}{33{,}6} + \frac{(30 - 36{,}4)^2}{36{,}4} + \frac{(32 - 38{,}4)^2}{38{,}4} + \frac{(48 - 41{,}6)^2}{41{,}6} = 4{,}396$$

Далее определяем число степеней свободы по формуле:

$$f = (r - 1) \times (c - 1)$$

Соответственно, для четырехпольной таблицы, в которой 2 ряда (r = 2) и 2 столбца (c = 2), число степеней свободы составляет:

$$f_{2x2} = (2 - 1) \times (2 - 1) = 1$$

Находим по таблице критическое значение критерия хи-квадрат Пирсона, которое при уровне значимости p=0,05 и числе степеней свободы 1 составляет 3,841.

Сравниваем значение критерия χ^2 с критическим значением при числе степеней свободы f (по таблице): 4,396>3,841, следовательно, зависимость частоты случаев артериальной гипертонии от наличия курения – статистически значима. Уровень значимости данной взаимосвязи соответствует p<0,05.

Таблица критических значений критерия χ^2 *Пирсона*, показана в табл. 5.

Таблица 5 – Критические значений χ^2 Пирсона

Число степеней свободы, f	χ^2 при p=0,05	χ^2 при p=0,01
1	3,841	6,635
2	5,991	9,21
3	7,815	11,345
4	9,488	13,277
5	11,07	15,086
6	12,592	16,812
7	14,067	18,475
8	15,507	20,09
9	16,919	21,666
10	18,307	23,209
11	19,675	24,725
12	21,026	26,217
13	22,362	27,688
14	23,685	29,141
15	24,996	30,578
16	26,296	32
17	27,587	33,409
18	28,869	34,805
19	30,144	36,191
20	31,41	37,566

Проведена работа с понятием «распределение "хи-квадрат"». Достигнута поставленная цель, а именно применение критерия на практике. А также были выполнены такие задачи, как предоставление теоретической части по теме и анализ применения распределения «хи-квадрат» в задачах статистического анализа данных.

Список использованных источников

1. Герасимович, А. И. Математическая статистика / А. И. Герасимович, Я. И. Матвеева. – Минск : Выш. школа, 1978. – 200 с.

2. Критерий хи-квадрат Пирсона. [Электронный ресурс]. – Режим доступа: http://medstatistic.ru/theory/hi_kvadrat.html. – Дата доступа: 25.03.2019.