УДК 519.2:330.43

08.00.13 Математические и инструментальные методы экономики (экономические науки)

**СТАТИСТИЧЕСКИЙ АНАЛИЗ ТАБЛИЦ ЧЕТЫРЕХ ПОЛЕЙ**

Муравьева Виктория Сергеевна
к.э.н., доцент

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор
РИНЦ SPIN-код: 4342-4994
*Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru*

Таблицу четырех полей строят для описания совместного распределения двух альтернативных признаков. Статья посвящена рассмотрению методов статистического анализа данных таблицы четырех полей в соответствии с традициями отечественной школы теории вероятностей и математической статистики. Такой анализ должен начинаться с выбора модели порождения данных. Применяют мультиномиальную модель и модель двух выборок. Проверяемые гипотезы и правила принятия решений меняются при переходе от одной модели к другой. Нельзя обоснованно судить о наличии связи между признаками только по величине тех или иных коэффициентов. Необходимо применять теорию проверки статистических гипотез. В мультиномиальной модели проверяют гипотезу независимости, а в модели двух выборок - гипотезу однородности долей. Только при отклонении нулевой гипотезы можно говорить о наличии связи между признаками, соответственно, о наличии эффекта при переходе от одной выборки к другой. Применяем метод вычисления асимптотических распределений функций от чисел в клетках таблицы четырех полей, основанный на многомерной центральной предельной теореме и методе линеаризации функций. Проверка статистических гипотез основана на использовании дисперсий коэффициентов ассоциации, коллигации и контингенции в мультиномиальной модели и разности выборочных долей в модели двух выборок. В применении дисперсий проявляется преимущество нашего подхода по сравнению с распространенной традицией. Некорректна встречающаяся в публикациях фраза: "Считается, что если коэффициент ассоциации превосходит 0,5 и коэффициент контингенции больше 0,3, то это свидетельствует о существенной связи между признаками". Говорить о "существенной связи между признаками" можно говорить лишь тогда, когда отклонена гипотеза независимости.

UDC 519.2:330.43

08.00.13 Mathematical and instrumental methods of Economics (economic sciences)

**STATISTICAL ANALYSIS OF FOUR-CELL TABLES**

Muravyeva Victoria Sergeevna
Cand.Econ.Sci., associate professor

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
*Bauman Moscow State Technical University, Moscow, Russia*

A four-cell table is constructed to describe the joint distribution of two alternative features. The article is devoted to the consideration of methods of statistical analysis of data from a table of four cells in accordance with the traditions of the Russian school of probability theory and mathematical statistics. This analysis should begin with the selection of a data generation model. A multinomial model and a two-sample model are applied. Testable hypotheses and decision rules change as you move from one model to another. It is impossible to reasonably judge the presence of a connection between signs only by the value of certain coefficients. It is necessary to apply the theory of testing statistical hypotheses. In the multinomial model, the hypothesis of independence is tested, and in the model of two samples, the hypothesis of homogeneity of shares. Only if the null hypothesis is rejected can we talk about the presence of a connection between the signs, respectively, about the presence of an effect when moving from one sample to another. We apply the method for calculating the asymptotic distributions of functions of numbers in the cells of the table of four cells, based on the multidimensional central limit theorem and the method of linearization of functions. Statistical hypothesis testing is based on the use of variances of the coefficients of association, colligation, and contingency in the multinomial model and the difference in sample shares in the model of two samples. The advantage of our approach over the widespread tradition is manifested in the use of variances. The phrase used in publications is incorrect: "It is believed that if the association coefficient exceeds 0.5 and the contingency coefficient is greater than 0.3, then this indicates a significant relationship between the characteristics." One can speak of an "essential connection between features" only when the hypothesis of independence is rejected. The recommendations obtained are based on the asymptotic normality of the considered coefficients.

Полученные рекомендации основаны на асимптотической нормальности рассматриваемых коэффициентов. Если в клетках таблицы четырех полей стоят сравнительно небольшие числа, то вместо предельных соотношений целесообразно пользоваться таблицами для конечных объемов выборок или соответствующими компьютерными программами

If the cells of the table of four cells contain relatively small numbers, then instead of limiting ratios, it is advisable to use tables for finite sample sizes or appropriate computer programs

Ключевые слова: СТАТИСТИЧЕСКИЙ АНАЛИЗ, ТАБЛИЦА ЧЕТЫРЕХ ПОЛЕЙ, МОДЕЛЬ ПОРОЖДЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ, МУЛЬТИНОМИАЛЬНАЯ МОДЕЛЬ, МОДЕЛЬ ДВУХ НЕЗАВИСИМЫХ ВЫБОРОК, ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ, КОЭФФИЦИЕНТ АССОЦИАЦИИ, КОЭФФИЦИЕНТ КОЛЛИГАЦИИ, КОЭФФИЦИЕНТ КОНТИНГЕНЦИИ, ПРОВЕРКА РАВЕНСТВА ДОЛЕЙ, АСИМПТОТИЧЕСКАЯ НОРМАЛЬНОСТЬ

Keywords: STATISTICAL ANALYSIS, TABLE OF FOUR FIELDS, MODEL OF PRODUCTION OF STATISTICAL DATA, MULTINOMIAL MODEL, MODEL OF TWO INDEPENDENT SAMPLES, STATISTICAL HYPOTHESIS TESTING, ASSOCIATION COEFFICIENT, COLLIGATION COEFFICIENT, CONTINGENCY COEFFICIENT, EQUALITY TESTING OF PROPORTIONS, ASYMPROTIC NORMALITY

http://dx.doi.org/10.21515/1990-4665-174-022

## 1. Введение

Работа посвящена современным математико-статистическим методам анализа таблиц четырех полей. Её необходимость основана на наличии некорректностей и ошибок в широко распространенных литературных источниках, как печатных, так и электронных. Работа выполнена в традициях отечественной научной школы в области теории вероятностей и математической статистики, основанной академиком А.Н. Колмогоровым.

## 2. Основные понятия

Предисловие к учебнику [1] начинается с констатации: "Прикладная статистика – это наука о том, как обрабатывать данные". Под данными понимаем любой вид зарегистрированной информации.

Базовым элементом для анализа является статистическая единица. Их объединение дает статистическую совокупность. Таким образом, в терминах математики статистическая совокупность - это множество, элементами которого являются статистические единицы. Генеральная

совокупность - множество всех объектов, которые имеют качества, свойства, интересующие исследователя. Выборочная совокупность (выборка) - совокупность элементов генеральной совокупности, информация о которых имеется у исследователя. Генеральная совокупность - обычно теоретическая конструкция, о свойствах которой узнаем по выборке. С развитием информационно-коммуникационных технологий анализа "больших данных" появилась возможность непосредственного анализа некоторых генеральных совокупностей, включающих большое количество статистических единиц, т.е. без промежуточного этапа в виде отбора и анализа элементов выборки. Примером является анализ генеральной совокупности научных публикаций, включенных в Российский индекс научного цитирования.

В статистике признаком называется функция, определенная для единиц статистической совокупности. Следовательно, значением признака (для конкретной статистической единицы) является значение этой функции. По утвердившейся традиции разными словами называют одни и те же сущности: совокупность - множество, признак - функция.

Примерами значений признаков (функций) являются числа, градации (элементы некоторого конечного множества), вектора, объекты нечисловой природы, другие математические объекты [2]. Если градации упорядочены, то говорят о порядковых признаках, если не упорядочены – о номинальных признаках. Для признаков, принимающих два возможных значения, используют ряд терминов - альтернативные, дихотомические, бинарные признаки.

В прикладной статистике термин "выборка" используют в двух смыслах. Во-первых, выборка - это часть генеральной совокупности (см. выше). Во вторых, это набор $X_1, X_2, ..., X_n$ реализаций (т.е. значений для некоторого полностью определенного элементарного исхода) независимых одинаково распределенных случайных величин. Число реализаций *n*

называется объемом выборки. Во втором случае, особенно если случайные величины имеют непрерывные функции распределения, термин "генеральная совокупность" обычно не используют, поскольку она должна состоять из бесконечного числа элементов (соответственно бесконечному числу возможных значений случайных величин) и тем самым может рассматривается не как реальный, а как теоретический (математический, условный) объект, для которого используют термин "пространство элементарных событий".

Отметим, что случайная величина - это не число, а функция, определенная на пространстве элементарных событий. Числом может быть реализация случайной величины, т.е. значение функции для определенного элементарного события. Наблюдаем двойственность терминологии. В теории вероятностей и математической статистике часто говорят - рассмотрим выборку $X_1, X_2, ..., X_n$, понимая под этим термином конечную последовательность независимых одинаково случайных распределенных случайных величин. В прикладной статистике при обработке реальных данных выборкой $X_1, X_2, ..., X_n$ называют конечную последовательность чисел. Отмеченная двойственность иногда приводи т к недоразумениям.

Здесь и далее определения и термины используем в соответствии со справочником [3].

## 3. Две вероятностно-статистические модели порождения таблиц четырех полей

Рассмотрим два альтернативных признака *X* и *У*, определенных на одном вероятностном пространстве и принимающих значения 0 и 1. Статистические данные часто представляют в виде таблицы четырех полей (табл.1).

Таблица 1.

**Таблица четырёх полей**

| Значения признаков | $X = 0$ | $X = 1$ | Сумма |
|---|---|---|---|
| $У = 0$ | $a$ | $b$ | $a + b$ |
| $У = 1$ | $c$ | $d$ | $c + d$ |
| Сумма | $a + c$ | $b + d$ | $n$ |

Этими полями являются ячейки таблицы с числами $a$, $b$, $c$, $d$, равными численности групп, соответствующих комбинациям значений признаков. При этом общий объем данных есть $n = a + b + c + d$.

Пример такой таблицы приведен в табл.2, в которой представлены данные опроса, соответствующие признакам "Возраст" ($X = 0$, если возраст опрошенного до 45 лет, $X = 1$ в противном случае) и "Желание путешествовать" ($У = 0$, если опрошенный высказывает такое желание, и $У = 1$ в противном случае). Отметим, что в табл.2 использованы значения соответствующих градаций альтернативных признаков, а не их условные обозначения в виде условных цифр 0 и 1).

Таблица 2.

**Пример таблицы четырёх полей**

| Возраст / Желание путешествовать | До 45 лет | После 45 лет | Сумма |
|---|---|---|---|
| Есть | 300 | 150 | 450 |
| Нет | 100 | 380 | 480 |
| Сумма | 400 | 530 | 930 |

Статистический анализ данных должен быть основан на той или иной вероятностно-статистической модели [4]. Для таблицы четырех полей есть две принципиально разные модели - мультиномиальная модель и модель двух независимых выборок.

В мультиномиальной модели предполагается, что пары $(X_i, Y_i)$, $i = 1,2,...,n$, являются независимыми одинаково распределенными случайными векторами. Их общее распределение задается вероятностями

$$p(a) = P(X_i = 0,\ Y_i = 0),\ \ p(b) = P(X_i = 1,\ Y_i = 0),$$
$$p(c) = P(X_i = 0,\ Y_i = 1),\ \ p(d) = P(X_i = 1,\ Y_i = 1),\ \ i = 1,2,...,n.$$

Здесь вероятности $p(a)$, $p(b)$, $p(c)$, $p(d)$ положительны и меньше 1, их сумма равна 1, т.е. $p(a) + p(b) + p(c) + p(d) = 1$, в остальном произвольны. Таким образом, распределение четырехмерного вектора $(a,b,c,d)$ задается тремя независимыми параметрами.

Это распределение является мультиномиальным (см., например, [5, разд. 6.3]). Распределение вектора $(a,b,c,d)$ таково:

$$p(k,m,t,q) = P(a = k, b = m, c = t, d = q) = \frac{n!}{k!m!t!q!}\, p(a)^k\, p(b)^m p(c)^t\, p(d)^q,$$

где $k,m,t,q$ - любые неотрицательные целые числа такие, что $k + m + t + q = n$.

Для мультиномиальной модели все четыре суммы (по строкам и по столбцам) $a + b$, $c + d$, $a + c$, $b + d$ являются случайными величинами.

В модели двух независимых выборок, наоборот, суммы по строкам зафиксированы: $a + b = n_1$, $c + d = n_2$, где $n_1$ и $n_2$ - заданные натуральные числа (объемы выборок). Альтернативный вариант - суммы по столбцам зафиксированы - переходит в рассматриваемый при симметрии матрицы табл.1 относительно главной диагонали, поэтому нет необходимости его рассматривать.

Таблица четырех полей в модели двух выборок переходит в табл.3.

Таблица 3.

**Таблица четырёх полей в модели двух независимых выборок**

| Значения признаков | $X = 0$ | $X = 1$ | Сумма |
|---|---|---|---|
| $Y = 0$ | $a$ | $b = n_1 - a$ | $n_1$ |
| $Y = 1$ | $c$ | $d = n_2 - c$ | $n_2$ |
| Сумма | $a + c$ | $b + d = n - a - c$ | $n$ |

В модели двух выборок, в отличие от мультиномиальной модели, пары $(X_i, Y_i)$, $i = 1,2,...,n$, не являются независимыми одинаково распределенными случайными векторами. Их общее распределение задается вероятностями

$$p(a) = P(X_i = 0,\ Y_i = 0),\ \ p(b) = P(X_i = 1,\ Y_i = 0) = 1 - p(a),$$
$$p(c) = P(X_i = 0,\ Y_i = 1),\ \ p(d) = P(X_i = 1,\ Y_i = 1) = 1 - p(c),\ \ i = 1,2,...,n.$$

Здесь вероятности $p(a), p(c)$ положительны и меньше 1, т.е. в остальном произвольны. Таким образом, распределение четырехмерного вектора $(a,b,c,d)$ задается двумя независимыми параметрами.

Случайная величина $a$ имеет биномиальное распределение $B(n_1, p(a))$, в то время как случайная величина $c$ также имеет биномиальное распределение $B(n_2, p(c))$, но, вообще говоря, с другими параметрами-значениями объема выборки и вероятности. Случайные величины $a$ и $c$ независимы. Следовательно, распределение вектора $(a,b,c,d)$ таково:

$$p(k,m,t,q) = P(a = k, b = m, c = t, d = q) = \frac{n_1! n_2!}{k!(n_1 - k)! t!(n_2 - t)!} p(a)^k (1 - p(a))^{n_1 - k} p(c)^t (1 - p(c))^{n_2 - t},$$

если $m = n_1 - k$, $q = n_2 - t$, и $p(k,m,t,q) = P(a = k, b = m, c = t, d = q) = 0$ в противном случае, где $k,m,t,q$ - любые неотрицательные целые числа такие, что $k \le n_1, m \le n_1, t \le n_2, q \le n_2$.

Поскольку распределения вектора $(a,b,c,d)$ в мультиномиальной модели и в модели двух выборок существенно отличаются, то и статистические выводы зависят от того, какая из двух моделей принята за основу.

Табл. 2 не дает ответа на вопрос о том, какая вероятностно-статистическая модель порождения данных использована. Эта таблица могла быть получена при двух различных схемах сбора данных.

В первой схеме опрашивают $n$ = 930 лиц. Предполагается, что они образуют представительную выборку из рассматриваемой генеральной

совокупности. Тогда анализ таблицы четырех полей следует проводить на основе мультиномиальной модели.

Во второй схеме заранее выделены две генеральные совокупности. В первую входят те, у кого есть желание путешествовать, во вторую - те, у кого нет такого желания. Из первой совокупности берется представительная выборка объема $n_1 = 450$, из второй - представительная выборка объема $n_2 = 480$. Тогда анализ таблицы четырех полей следует проводить на основе модели двух независимых выборок.

## 4. Анализ таблицы четырех полей для мультиномиальной модели

Для упрощения дальнейшего изложения целесообразно несколько изменить обозначения.

В вероятностной модели $X = X(w)$ и $Y = Y(w)$ - случайные величины, принимающие два значения - 0 и 1. Здесь $w$ - элемент пространства элементарных исходов. Пусть $p_1 = P(X(w) = 1)$ и $p_2 = P(Y(w) = 1)$. Вероятности получения чисел в ячейках таблицы четырех полей четырьмя числами:

$$p(a) = P(X(w) = 0, Y(w) = 0) = p_{00}, p(b) = P(X(w) = 1, Y(w) = 0) = p_{10},$$

$$p(c) = P(X(w) = 0, Y(w) = 1) = p_{01}, \ p(d) = P(X(w) = 1, Y(w) = 1) = p_{11.}$$

Очевидно, верны равенства:

$$p_{00} + p_{10} + p_{01} + p_{11} = 1, \ \ p_{10} + p_{11} = p_1, \ p_{01} + p_{11} = p_2.$$

В табл.4 сведены вместе введенные выше вероятности.

Таблица 4.

**Вероятности в мультиномиальной модели**

| Значения признаков | $X = 0$ | $X = 1$ | Всего |
|---|---|---|---|
| $Y = 0$ | $p_{00}$ | $p_{10}$ | $1 - p_2$ |
| $Y = 1$ | $p_{01}$ | $p_{11}$ | $p_2$ |
| Всего | $1 - p_1$ | $p_1$ | 1 |

Обычно выделяют три важных частных случая - поглощения, несовместности и независимости признаков. Другими словами, поглощения, несовместности и независимости событий $\{w: X(w) = 1\}$ и $\{w: Y(w) = 1\}$.

В случае поглощения одно из этих событий содержит другое, а потому

$$p_{00} = 1 - \max(p_1 , p_2).$$

Если событие $\{w: X(w) = 1\}$ содержит событие $\{w: Y(w) = 1\}$, то из $Y(w) = 1$ следует, что $X(w) = 1$, т.е. событие $\{w: X(w) = 0, Y(w) = 1\}$ невозможно. В этом случае $p(c) = P(X(w) = 0, Y(w) = 1) = 0$, а потому $c = 0$ с вероятностью 1. Обратное неверно - из $c = 0$ не следует, что обязательно имеем случай поглощения. Но такое предположение напрашивается.

Если, наоборот, событие $\{w: Y(w) = 1\}$ содержит событие $\{w: X(w) = 1\}$, то из $X(w) = 1$ следует, что $Y(w) = 1$, т.е. событие $\{w: X(w) = 1, Y(w) = 0\}$ невозможно. В этом случае $p(b) = P(X(w) = 1, Y(w) = 0) = 0$, а потому $b = 0$ с вероятностью 1. Обратное неверно - из $b = 0$ не следует, что обязательно имеем случай поглощения. Но такое предположение напрашивается.

В случае несовместности

$$p_{00} = 1 - p_1 - p_2.$$

Несовместность событий $\{w: X(w) = 1\}$ и $\{w: Y(w) = 1\}$ означает, что событие $\{w: X(w) = 1, Y(w) = 1\}$ невозможно, $p(d) = P(X(w) = 1, Y(w) = 1) = p_{11} = 0$, а потому $d = 0$ с вероятностью 1. Обратное неверно - из $d = 0$ не следует, что обязательно имеем случай несовместности. Но такое предположение напрашивается.

Независимость признаков $X$ и $Y$ - это справедливость равенств

$$P(X(w) = 0, Y(w) = 0) = P(X(w) = 0)P(Y(w) = 0),$$
$$P(X(w) = 1, Y(w) = 0) = P(X(w) = 1)P(Y(w) = 0),$$
$$P(X(w) = 0, Y(w) = 1) = P(X(w) = 0)P(Y(w) = 1),$$
$$P(X(w) = 1, Y(w) = 1) = P(X(w) = 1)P(Y(w) = 1),$$

(по определению независимости случайных величин в теории вероятностей [3]). Нетрудно проверить, что все эти равенства вытекают из первого из них. Поэтому независимость признаков имеет место тогда и только тогда, когда

$$p_{00} = (1 - p_1)(1 - p_2) = 1 - p_1 - p_2 + p_1p_2. \quad (1)$$

Пусть исходные данные - таблица четырех полей, полученная в предположениях мультиномиальной модели. Для проверки признаков $X$ и $Y$, т.е. равенства (1), следует применить теорию проверки статистических гипотез.

Нулевая гипотеза имеет вид:

$$H_0 : p_{00} = 1 - p_1 - p_2 + p_1p_2. \quad (2)$$

В качестве альтернативной гипотезы $H_1$ рассмотрим отрицание нулевой гипотезы $H_0$:

$$H_1 : p_{00} \neq 1 - p_1 - p_2 + p_1p_2. \quad (3)$$

(в некоторых прикладных задачах могут быть полезны другие альтернативные гипотезы, например, полученные из (3) заменой $\neq$ на $<$ или $>$ ; от выбора альтернативной гипотезы зависит вид критической области [3]).

С помощью равносильных преобразований формулам (1) и (2) можно придать другой вид. Можно говорить о статистической проверке нулевой гипотезы

$$H_0: p_{11} = p_1 p_2 \quad (4)$$

(что эквивалентно проверке равенства $p_{00} = (1 – p_1)(1 – p_2)$).

Преобразуем равенство (4):

$$p_{11} = (p_{10} + p_{11})(p_{01} + p_{11}) \quad (5)$$

(см. табл.4). Поскольку сумма всех вероятностей попадания в ячейки таблицы четырех полей равно 1, то из (5) следует, что

$$p_{11} (p_{00} + p_{01} + p_{10} + p_{11}) = (p_{10} + p_{11})(p_{01} + p_{11}) \quad (6)$$

Раскроем скобки в обеих частях соотношения (6):

$$p_{11}\,p_{00} + p_{11}p_{01} + p_{11}p_{10} + p_{11}p_{11} = p_{10}\,p_{01} + p_{11}\,p_{01} + p_{10}\,p_{11} + p_{11}\,p_{11}. \quad (7)$$

Сокращая равные слагаемые в левой и правой частях равенства (7), получаем, что

$$p_{11}\,p_{00} = p_{10}\,p_{01}. \quad (8)$$

Следовательно, гипотеза о справедливости равенства (4) эквивалентна гипотезе

$$H_0 : p_{00}\,p_{11} - p_{10}\,p_{01} = 0 \quad (9)$$

при альтернативной гипотезе

$$H_1 : p_{00}\,p_{11} - p_{10}\,p_{01} \neq 0. \quad (10)$$

Как уже отмечалось, четырехмерный случайный вектор ($a$, $b$, $c$, $d$) (см. табл.1) имеет мультиномиальное распределение с числом испытаний $n$ и вектором вероятностей исходов ($p_{00}$, $p_{10}$, $p_{01}$, $p_{11}$). Как следует из многомерного закона больших чисел [3], состоятельными оценками этих вероятностей являются дроби $a/n$, $b/n$, $c/n$, $d/n$ соответственно, т.е .

$$\lim_{n\to\infty}\left(\frac{a}{n},\frac{b}{n},\frac{c}{n},\frac{d}{n}\right) = (p_{00}, p_{10}, p_{01}, p_{11}) \quad (11)$$

(сходимость по вероятности $d$). Следовательно, критерий проверки гипотезы (9) может быть основан на статистике

$$Z = ad - bc, \quad (12)$$

поскольку из (10) и теоремы о наследовании сходимости [1] вытекает, что при справедливости этой гипотезы

$$\lim_{n\to\infty}\left(\frac{1}{n^2}Z\right) = p_{00}\,p_{11} - p_{10}\,p_{01} = 0, \quad (13)$$

а при альтернативной гипотезе этот предел не равен 0.

### 5. Асимптотическое распределение статистики *Z*

С целью проверки гипотезы (9) найдем асимптотическое распределение статистики $Z$.

Начнем с асимптотического распределения вектора случайного вектора (*a, b, c, d*). Согласно многомерной центральной предельной теореме (см., например, [1, 3]) вектор

$$\xi = \sqrt{n}\left(\frac{a}{n} - p_{00}, \frac{b}{n} - p_{10}, \frac{c}{n} - p_{01}, \frac{d}{n} - p_{11}\right) \tag{14}$$

имеет в асимптотике четырехмерное нормальное распределение с математическим ожиданием (0, 0, 0, 0) (в соответствии с (11)) и ковариационной матрицей

$$\operatorname{cov}(\xi) = \begin{pmatrix} p_{00}(1-p_{00}) & -p_{00}p_{10} & -p_{00}p_{01} & -p_{00}p_{11} \\ -p_{00}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{01} & -p_{10}p_{11} \\ -p_{00}p_{01} & -p_{10}p_{01} & p_{01}(1-p_{01}) & -p_{01}p_{11} \\ -p_{00}p_{11} & -p_{10}p_{11} & -p_{01}p_{11} & p_{11}(1-p_{11}) \end{pmatrix} \tag{15}$$

(см., например, в [5, с.153] формулу (6.3.5) для смешанных моментов мультиномиального распределения). Следовательно, ковариационная матрица для частот имеет вид

$$\operatorname{cov}\left(\frac{a}{n}, \frac{b}{n}, \frac{c}{n}, \frac{d}{n}\right) = \frac{1}{n}\operatorname{cov}(\xi). \tag{16}$$

Для нахождения асимптотического распределения статистики *Z* применим метод линеаризации (см., например, [1]). В силу (11) для любой достаточно гладкой функции *g*(*x, y, t, w*) имеем:

$$\begin{aligned} &g\left(\frac{a}{n}, \frac{b}{n}, \frac{c}{n}, \frac{d}{n}\right) - g(p_{00}, p_{10}, p_{01}, p_{11}) = \\ &\frac{\partial g}{\partial x}\left(\frac{a}{n} - p_{00}\right) + \frac{\partial g}{\partial y}\left(\frac{b}{n} - p_{10}\right) + \frac{\partial g}{\partial t}\left(\frac{c}{n} - p_{01}\right) + \frac{\partial g}{\partial w}\left(\frac{d}{n} - p_{11}\right) + \ldots \end{aligned} \tag{17}$$

где частные производные берутся в точке $(p_{00}, p_{10}, p_{01}, p_{11})$, а многоточием обозначены бесконечно малые более высокого порядка, чем бесконечно малые во второй строке формулы (16). Следовательно, асимптотическое распределение приращения функции *g*(*x, y, t, w*) (первая строка формулы (16)) определяется главным линейным членом (вторая строка формулы (16)).

Введем в рассмотрение функцию

$$g(x, y, t, w) = xw - yt. \quad (18)$$

Её частные производные в точке $(x, y, t, w)$ таковы:

$$\frac{\partial g}{\partial x} = w, \frac{\partial g}{\partial y} = -t, \frac{\partial g}{\partial t} = -y, \frac{\partial g}{\partial w} = x. \quad (19)$$

При справедливости нулевой гипотезы (9 $g(p_{00}, p_{10}, p_{01}, p_{11}) = 0$.

Из (17) следует, что в точке $(x, y, t, w) = (p_{00}, p_{10}, p_{01}, p_{11})$ асимптотическая дисперсия такова:

$$\begin{aligned} &D(g\left(\frac{a}{n}, \frac{b}{n}, \frac{c}{n}, \frac{d}{n}\right) - g(p_{00}, p_{10}, p_{01}, p_{11})) = \\ &= D\left(p_{11}\left(\frac{a}{n} - p_{00}\right) - p_{01}\left(\frac{b}{n} - p_{10}\right) - p_{10}\left(\frac{c}{n} - p_{01}\right) + p_{00}\left(\frac{d}{n} - p_{11}\right)\right). \end{aligned} \quad (20)$$

Поскольку

$$M\left(\frac{a}{n} - p_{00}\right) = M\left(\frac{b}{n} - p_{10}\right) = M\left(\frac{c}{n} - p_{01}\right) = M\left(\frac{d}{n} - p_{11}\right) = 0, \quad (21)$$

то, воспользовавшись (21), получаем:

$$\begin{aligned} &D\left(p_{11}\left(\frac{a}{n} - p_{00}\right) - p_{01}\left(\frac{b}{n} - p_{10}\right) - p_{10}\left(\frac{c}{n} - p_{01}\right) + p_{00}\left(\frac{d}{n} - p_{11}\right)\right) = \\ &= M\left(p_{11}\left(\frac{a}{n} - p_{00}\right) - p_{01}\left(\frac{b}{n} - p_{10}\right) - p_{10}\left(\frac{c}{n} - p_{01}\right) + p_{00}\left(\frac{d}{n} - p_{11}\right)\right)^2 = \frac{1}{n^2}Q, \end{aligned}$$

где

$$\begin{aligned} Q = {} & p_{11}^2 D(a) + p_{01}^2 D(b) + p_{10}^2 D(c) + p_{00}^2 D(d) - 2p_{11}p_{01}\operatorname{cov}(a,b) - 2p_{11}p_{10}\operatorname{cov}(a,c) + \\ & + 2p_{11}p_{00}\operatorname{cov}(a,d) + 2p_{01}p_{10}\operatorname{cov}(b,c) - 2p_{01}p_{00}\operatorname{cov}(b,d) - 2p_{10}p_{00}\operatorname{cov}(c,d). \end{aligned}$$

Далее, подставив значения дисперсий и ковариаций из (15), получаем, что

$$D(g\left(\frac{a}{n}, \frac{b}{n}, \frac{c}{n}, \frac{d}{n}\right) = \frac{1}{n}A,$$

где

$$\begin{aligned} A = {} & p_{11}^2 p_{00}(1 - p_{00}) + p_{01}^2 p_{10}(1 - p_{10}) + p_{10}^2 p_{01}(1 - p_{01}) + p_{00}^2 p_{11}(1 - p_{11}) + \\ & + 2p_{11}p_{01}p_{00}p_{10} + 2p_{11}p_{10}p_{00}p_{01} - 2p_{11}p_{00}p_{00}p_{11} - 2p_{01}p_{10}p_{00}p_{01} + \\ & + 2p_{01}p_{00}p_{10}p_{11} + 2p_{10}p_{00}p_{01}p_{11}. \end{aligned}$$

Приведем подобные члены:

$$A = p_{11}^2 p_{00} + p_{01}^2 p_{10} + p_{10}^2 p_{01} + p_{00}^2 p_{11} - 4p_{11}^2 p_{00}^2 - 2p_{01}^2 p_{10}^2 + 6p_{11}p_{01}p_{00}p_{10}. \quad (22)$$

Для получения выборочной оценки $A^*$ величины достаточно заменить в формуле (22) вероятности $p_{00}, p_{10}, p_{01}, p_{11}$ на частоты, т.е. на $\frac{a}{n}, \frac{b}{n}, \frac{c}{n}, \frac{d}{n}$ соответственно.

Итак, при справедливости гипотезы независимости (9) асимптотическое распределение величины $n^{-2}Z$ - нормальное с нулевым математических ожиданием (см.(13)) и выборочной оценкой дисперсии $A^*/n$ (см. (22)). Следовательно, асимптотическое правило принятия решения при проверке гипотезы независимости (9) на уровне значимости $\alpha$ таково: если

$$\frac{|Z|}{n^2} \le C(\alpha)\sqrt{\frac{A^*}{n}}, \ m.e. \ |Z| \le C(\alpha) n\sqrt{n}\sqrt{A^*},$$

то гипотезу независимости (9) принять, в противном случае отклонить. Как обычно при использовании асимптотически нормальных критериев проверки статистических гипотез,

$$C(\alpha) = \Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

где $\Phi^{-1}$ - функция, обратная к функции стандартного нормального распределения [1]. Для наиболее распространенного при анализе статистических данных значения $\alpha = 0{,}05$ используем коэффициент $C(0{,}05) = 1{,}96$.

## 6. Коэффициенты связи альтернативных признаков

Вслед за М.Дж. Кендаллом и А. Стьюартом [6, гл.33] рассмотрим коэффициенты ассоциации и коллигации Юла [7, 8] и коэффициент контингенции Пирсона [9].

Как измерять величину связи между двумя признаками одним коэффициентом - мерой связи? С помощью какого коэффициента это делать? Естественно потребовать, чтобы были известны границы изменения этого коэффициента, а также, чтобы он принимал выделенное

значение - среднее или нижнее - в интервале изменения, если признаки не связаны (т.е. независимы). Выбирая начало отсчета и единицу измерения, можно любой такой коэффициент заставить изменяться в интервале [-1, +1], причем случай независимости будет соответствовать нулевому значению коэффициента. Это удобно тем, что свойства рассматриваемого коэффициента напоминают свойства коэффициентов корреляции К. Пирсона и Спирмена.

Какие значения может принимать коэффициент $Z = ad - bc$ (см. (12))? Для ответа на этот вопрос надо решить две задачи оптимизации:

$$\begin{cases} ad - bc \to \max \\ a + b + c + d = n, \\ a \ge 0, b \ge 0, c \ge 0, d \ge 0, \end{cases} \quad \begin{cases} ad - bc \to \min \\ a + b + c + d = n, \\ a \ge 0, b \ge 0, c \ge 0, d \ge 0 . \end{cases}$$

В первой из них необходимо минимизировать $bc$, при этом во второй строке минимизировать $b + c$. Следовательно, надо положить $b = c = 0$. Тогда $d = n - a$ и остается максимизировать $a(n-a)$. Как известно, максимум достигается при $a = n/2$ (для четного $n$) и равен $n^2/4$. Аналогично во второй задаче необходимо минимизировать $ad$, при этом во второй строке минимизировать $a + d$. Следовательно, надо положить $a = d = 0$. Тогда $c = n - b$ и остается минимизировать $(-b(n-b))$. Как известно, максимум достигается при $b = n/2$ и равен $(-n^2/4)$. Для нечетного $n$ увеличивается количество точек, в которых достигается максимум или минимум коэффициента $Z$.

Следовательно, коэффициент $Z$ принимает значения от $(-n^2/4)$ до $(+n^2/4)$. Этот коэффициент не удовлетворяет сформулированному выше требованию, а потому он не может непосредственно применяться для оценки связи между альтернативными признаками. Нормированный коэффициент $Z_0 = \frac{4}{n^2} Z$ удовлетворяет сформулированному выше условию и может использоваться в качестве меры связи двух альтернативных

признаков. Асимптотическое распределение этого коэффициента найдено в предыдущем разделе 5 настоящей статьи, там же разобраны правила проверки статистической гипотезы независимости.

Однако крайнее значение (-1) достигается (при четном $n$) лишь в одной точке - при $a = d = 0$ и $b = c = n/2$, а крайнее значение (+1) тоже лишь в одной точке - при $b = c = 0$ и $a = d = n/2$. Точки, в которых достигаются крайние значения, не связаны с интуитивным представлением о независимости. При нечетном $n$ экстремальных точек больше, но они не связаны с представлением о независимости. Поэтому важно рассмотреть другие коэффициенты, измеряющие связь между двумя альтернативными признаками.

### 7. Коэффициент ассоциации

*Коэффициент ассоциации* введен и изучен английским статистиком Джорджем Одни Юлом (1871-1951) в статьях [7, 8] и ныне носит его имя:

$$Q = \frac{ad - bc}{ad + bc} = \frac{Z}{ad + bc}. \qquad (23)$$

Этот коэффициент близок к 0, если признаки независимы (при безграничном росте объема данных возможна замена частот вероятностями, и тогда Q = 0 соответствует независимости признаков в мультиномиальной модели). Очевидно, $Q = 1$ тогда и только тогда, когда $bc = 0$, и $Q = (-1)$ тогда и только тогда, когда $ad = 0$, следовательно, равенство $|Q| = 1$ соответствует (в асимптотике) условию поглощения.

При справедливости гипотезы (9) независимости признаков и безграничном росте объема $n$ данных коэффициент контингенции $Q$ является асимптотически нормальной случайной величиной с нулевым математическим ожиданием и дисперсией $D(Q)$, выборочной оценкой которой является $D^*(Q)$ следующего вида:

$$D^*(Q) = \frac{1}{4}(1-Q^2)^2\left\{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right\} \quad (24)$$

(см. [6, п.33.8]). Следовательно, асимптотическое правило принятия решения при проверке гипотезы независимости (9) на уровне значимости $\alpha$ таково: если

$$|Q| \le C(\alpha)\sqrt{D^*(Q)}, \quad (25)$$

то гипотезу независимости (9) принять, в противном случае отклонить. Как обычно при использовании асимптотически нормальных критериев проверки статистических гипотез,

$$C(\alpha) = \Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

где $\Phi^{-1}$ - функция, обратная к функции стандартного нормального распределения [1]. Для наиболее распространенного при анализе статистических данных значения $\alpha = 0{,}05$ используем коэффициент $C(0{,}05) = 1{,}96$.

Метод получения асимптотического распределения статистики $Q$ - тот же, что и для получения асимптотического распределения статистики $Z$ выше в п.5 настоящей статьи, поэтому выкладки не приводим (см. статьи Юла [7, 8]). Отметим, что согласно (24) асимптотическая дисперсия меняется пропорционально $1/n$ (поскольку величины в фигурных скобках убывают как $1/n$), правая часть неравенства (25) убывает пропорционально $1/\sqrt{n}$, следовательно, область принятия гипотезы независимости сужается.

*Пример 1.* Для данных табл. 2 имеем:

$$Q = \frac{300\times380-150\times100}{300\times380+150\times100} = \frac{114000-15000}{114000+15000} = \frac{99}{129} = 0{,}767$$

и

$$D^*(Q) = \frac{1}{4}(1-0{,}767^2)^2\left\{\frac{1}{300}+\frac{1}{150}+\frac{1}{100}+\frac{1}{380}\right\} =$$

$$= 0{,}25\times0{,}170\{0{,}00333+0{,}00667+0{,}01+0{,}00263\} = 0{,}0424\times0{,}02263 = 0{,}00096.$$

Следовательно, правая часть неравенства (25) равна

$$1{,}96\sqrt{0{,}00096} = 1{,}96 \times 0{,}0310 = 0{,}0607$$

Поскольку 0,767 > 0,0607, гипотеза независимости отклоняется.

**8. Коэффициент коллигации**

Второй коэффициент, рассмотренный Юлом в [7, 8] (см. [6, с. 723]), называется *коэффициентом коллигации* и имеет вид:

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}. \qquad (26)$$

*Пример 2.* Для данных табл. 2 имеем:

$$Y = \frac{\sqrt{300 \times 380} - \sqrt{150 \times 100}}{\sqrt{300 \times 380} + \sqrt{150 \times 100}} = \frac{\sqrt{114000} - \sqrt{15000}}{\sqrt{114000} + \sqrt{15000}} = \frac{337{,}64 - 122{,}47}{337{,}64 + 122{,}47} = \frac{215{,}17}{460{,}11} = 0{,}468$$

Согласно [6] нетрудно показать, что

$$Q = \frac{2Y}{1 + Y^2}$$

Действительно, для данных табл. 2 имеем:

$$\frac{2Y}{1 + Y^2} = \frac{2 \times 0{,}468}{1 + 0{,}468^2} = \frac{0{,}936}{1{,}219} = 0{,}768 = Q$$

(с точностью до 0,001, расхождение из-за погрешностей вычислений).

При справедливости гипотезы (9) независимости признаков и безграничном росте объема $n$ данных коэффициент коллигации $Y$ является асимптотически нормальной случайной величиной с нулевым математическим ожиданием и дисперсией $D(Y)$, выборочной оценкой которой является $D^*(Y)$ следующего вида:

$$D^*(Y) = \frac{1}{16}(1 - Y^2)^2 \left\{ \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right\} \qquad (27)$$

(см. [6, п.33.8]). Как и для коэффициента ассоциации, асимптотическое правило принятия решения при проверке гипотезы независимости (9) на уровне значимости $\alpha$ таково: если

$$|Y| \le C(\alpha)\sqrt{D^*(Y)}, \qquad (28)$$

то гипотезу независимости (9) принять, в противном случае отклонить. Как и ранее, здесь

$$C(\alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

где $\Phi^{-1}$ - функция, обратная к функции стандартного нормального распределения [1]. Наиболее распространенному при анализе статистических данных значению уровня значимости $\alpha$ = 0,05 соответствует коэффициент $C(0,05) = 1,96$.

Метод получения асимптотического распределения статистики $Y$ - тот же, что и для получения асимптотического распределения статистики $Z$ в п.5 настоящей статьи и статистики $Q$ выше, поэтому выкладки не приводим (см. статьи Юла [7, 8]). Отметим, что согласно (27) асимптотическая дисперсия меняется пропорционально $1/n$ (поскольку величины в фигурных скобках убывают как $1/n$), правая часть неравенства (28) убывает пропорционально $1/\sqrt{n}$, следовательно, область принятия гипотезы независимости сужается.

*Пример 3.* Для данных табл. 2 имеем:

$$D^*(Y) = \frac{1}{16}(1 - 0,468^2)^2\left\{\frac{1}{300} + \frac{1}{150} + \frac{1}{100} + \frac{1}{380}\right\} = 0,0381 \times 0,02263 = 0,000862.$$

Для уровня значимости $\alpha$ = 0,05 правая часть неравенства (28) равна

$$1,96\sqrt{0,0011} = 1,96 \times 0,02936 = 0,0575.$$

Поскольку 0,468 > 0,0575, гипотеза (9) независимости признаков отклоняется.

### 9. Коэффициент контингенции

Используют и третий коэффициент связи между альтернативными признаками - *коэффициент контингенции*, изученный Э. Пирсоном (1895-1980) в статье [9]):

$$V = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \quad (29)$$

*Пример 4.* Для данных табл. 2 имеем:

$$V = \frac{300\times380 - 150\times100}{\sqrt{(300+150)(300+100)(150+380)(100+380)}} = \frac{114000-15000}{\sqrt{450\times400\times530\times480)}} =$$
$$= \frac{99000}{\sqrt{45792000000}} = \frac{99}{\sqrt{45792}} = \frac{99}{213{,}99} = 0{,}462.$$

При справедливости гипотезы (9) независимости признаков и безграничном росте объема $n$ данных коэффициент контингенции $V$ является асимптотически нормальной случайной величиной с нулевым математическим ожиданием и дисперсией $D(V)$, выборочной оценкой которой является $D^*(V)$ следующего вида:

$$D^*(V) = \frac{1}{n}\left(1 - V^2\right) + \frac{1}{n}\left(V + \frac{1}{2}V^2\right)\frac{(a-d)^2 - (b-c)^2}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} -$$
$$- \frac{3}{4n}V^2\left\{\frac{(a+b-c-d)^2}{(a+b)(c+d)} - \frac{(a+c-b-d)^2}{(a+c)(b+d)}\right\} \quad (30)$$

(см. [6, п.33.8]). Как и для коэффициентов ассоциации и коллигации, асимптотическое правило принятия решения при проверке гипотезы независимости (9) на уровне значимости $\alpha$ таково: если

$$|V| \le C(\alpha)\sqrt{D^*(V)}, \quad (31)$$

то гипотезу независимости (9) принять, в противном случае отклонить. Как и ранее, здесь

$$C(\alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

где $\Phi^{-1}$ - функция, обратная к функции стандартного нормального распределения [1]. Наиболее распространенному при анализе статистических данных значению уровня значимости $\alpha$ = 0,05 соответствует коэффициент $C(0{,}05) = 1{,}96$.

Метод получения асимптотического распределения статистики $V$ - тот же, что и для получения асимптотического распределения статистики $Z$ в п.5 настоящей статьи и статистик $Q$, $Y$ выше, поэтому выкладки не

приводим (см. работу Э. Пирсона [9]). Отметим, что согласно (30) асимптотическая дисперсия меняется пропорционально 1/*n*, следовательно, правая часть неравенства (31) убывает пропорционально $1/\sqrt{n}$, а потому область принятия гипотезы независимости сужается.

*Пример 4.* Для данных табл. 2 имеем:

$$D^*(V)=\frac{1}{930}\left(1-0{,}462^2\right)+\frac{1}{930}\left(0{,}462+\frac{1}{2}0{,}462^2\right)\frac{(300-380)^2-(150-100)^2}{\sqrt{(300+150)(300+100)(150+380)(100+380)}}-$$

$$-\frac{3\times 0{,}462^2}{4\times 930}\left\{\frac{(300+150-100-380)^2}{(300+150)(100+380)}-\frac{(300+100-150-380)^2}{(300+100)(150+380)}\right\}=$$

$$=\frac{1}{930}\left[0{,}7866+0{,}5687\times\frac{6400-2500}{213990}-0{,}1601\left\{\frac{900}{216000}-\frac{16900}{212000}\right\}\right]=$$

$$=\frac{1}{930}\left[0{,}7866+0{,}5687\times 0{,}01823-0{,}1601\{0{,}00417-0{,}0797\}\right]=\frac{1}{930}\left[0{,}7866+0{,}0104+0{,}0121\right]=$$

$$=\frac{0{,}8091}{930}=0{,}00087.$$

Для уровня значимости $\alpha$ = 0,05 правая часть неравенства (31) равна

$$1{,}96\sqrt{0{,}00087}=1{,}96\times 0{,}0293=0{,}0574.$$

Поскольку 0,462 > 0,0574, гипотеза (9) независимости признаков отклоняется.

## 10. О свойствах коэффициентов ассоциации, коллигации и контингенции

Отметим, что все три коэффициента ассоциации, коллигации и контингенции принимают значения на отрезке [-1, +1], причем крайние значения достигаются.

В различных материалах, размещенных в Интернете, а также в печатных публикациях имеется некорректная фраза: "Считается, что если $K_{асс}$ > 0,5 и $K_{конт}$ > 0,3, это свидетельствует о существенной связи между признаками". Здесь в наших обозначениях $K_{асс}$ = *Q* и $K_{конт}$ =*V*. Приведенная фраза некорректна прежде всего потому, что игнорирует возможность получения отрицательных значений коэффициентов ассоциации и контингенции, достаточно больших по абсолютной

величине. Заменив в неравенствах значения коэффициентов на их абсолютные значения, т.е. перейдя к неравенствам $|K_{acc}| > 0,5$ и $|K_{конт}| > 0,3$, но остается другая. Говорить о "существенной связи между признаками" можно говорить лишь тогда, когда отклонена гипотеза независимости. Если в клетках таблицы четырех полей стоят сравнительно небольшие числа, то может случиться так, что хотя рассматриваемые коэффициенты заметно отличаются от 0, но тем не менее гипотеза независимости отклоняется.

*Пример 5.* Для данных табл. 5 коэффициент ассоциации Q = 0,5135, т.е. заметно отличается от 0, но при этом правая часть неравенства (25) равна 0,6538, и поскольку 0,5135 < 0,6538, гипотеза независимости принимается.

Таблица 5.

**Второй пример таблицы четырёх полей**

| Значения признаков | $X = 0$ | $X = 1$ | Сумма |
|---|---|---|---|
| $Y = 0$ | 14 | 3 | 17 |
| $Y = 1$ | 6 | 4 | 10 |
| Сумма | 20 | 7 | 27 |

Приведенные выше рассуждения основаны на асимптотической нормальности рассматриваемых коэффициентов. Если в клетках таблицы четырех полей стоят сравнительно небольшие числа, то вместо предельных соотношений целесообразно пользоваться результатами для конечных объемов выборок. Например, таблицами в классическом сборнике [10, табл. 5.6]. В этой книге, составленной членами-корреспондентами АН СССР Л.Н. Большевым и Н.В. Смирновым, представлены основные расчетные инструменты математической статистики XX в. В этой книге - не только подробные таблицы. Пояснительная часть книги представляет собой справочник по

статистическим и вычислительным методам, применяемым при решении задач математической статистики. В современных условиях вместо таблиц могут быть применены соответствующие компьютерные программы, с помощью которых, грубо говоря, могут быть рассчитаны разделы таблиц для конкретных статистических данных, в том числе содержащихся в интересующей исследователя таблице четырех полей.

Анализу таблиц сопряженности, частным случаем которых являются таблиц четырех полей, посвящена монография Г. Аптона [11]. Задачи статистического приемочного контроля, основанные на применении таблиц четырех полей, рассмотрены нами в статье [12].

## 11. Анализ таблицы четырех полей для модели двух независимых выборок

Примем, что в таблице четырех полей для модели двух независимых выборок заданы суммы по строкам (альтернативный вариант, в котором заданы суммы по столбцам, рассматривается аналогично). Итак, в этой модели суммы по строкам зафиксированы: $a + b = n_1$, $c + d = n_2$, где $n_1$ и $n_2$ - заданные натуральные числа (объемы выборок), как показано в табл. 6.

Таблица 6.

**Таблица четырёх полей в модели двух независимых выборок**

| Значения признаков | $X = 0$ | $X = 1$ | Сумма |
|---|---|---|---|
| $Y = 0$ | $a$ | $b = n_1 - a$ | $n_1$ |
| $Y = 1$ | $c$ | $d = n_2 - c$ | $n_2$ |
| Сумма | $a + c$ | $b + d = n - a - c$ | $n$ |

В модели двух независимых выборок имеются две независимые случайные величины $a$ и $c$. Каждая из них имеет биномиальное распределение, $B(n_1, q_1)$ и $B(n_2, q_2)$ соответственно, т.е.

$$P(a=k)=C_{n_1}^{k}q_1^{k}(1-q_1)^{n_1-k}, k=0,1,2,\ldots n_1,\ \ P(c=m)=C_{n_2}^{m}q_2^{m}(1-q_2)^{n_2-m}, m=0,1,2,\ldots n_2.$$

Методы доверительного оценивания параметров $q_1$ и $q_2$ представлены в [13, 14].

В рассматриваемой модели двух выборок в качестве центральной проблемы изучения выступает проверка статистической гипотезы о равенстве вероятностей $q_1$ и $q_2$, в отличие от гипотезы независимости признаков в мультиномиальной модели. Обсудим проверку нулевой гипотезы (гипотезы однородности долей)

$$H_0: q_1 = q_2 \quad (32)$$

при альтернативной гипотезе

$$H_1: q_1 \neq q_2,$$

являющейся отрицанием нулевой гипотезы и означающей наличие эффекта при переходе от одной выборки к другой. Гипотезу (32) называют также гипотезой однородности долей, чтобы отметить ее включение в систему моделей и методов проверки однородности двух независимых выборок [15 - 17].

Состоятельные несмещенные оценки вероятностей $q_1$ и $q_2$ таковы:

$$q_1^* = \frac{a}{n_1} = \frac{a}{a+b}, \; q_2^* = \frac{c}{n_2} = \frac{c}{c+d}.$$

Когда объемы выборок безгранично растут, $n_1 \to +\infty, n_2 \to +\infty,$ частоты сходятся к вероятностям, $q_1^* \to q_1, \; q_2^* \to q_2,$ а потому

$$q_1^* - q_2^* \to q_1 - q_2. \quad (33)$$

При справедливости нулевой гипотезы (32) правая часть соотношения (33) равна 0, а потому проверку (32) естественно проводить на основе величины

$$q_1^* - q_2^* = \frac{a}{a+b} - \frac{c}{c+d} = \frac{ac+ad-ac-bc}{(a+b)(c+d)} = \frac{ad-bc}{(a+b)(c+d)} = \frac{Z}{(a+b)(c+d)},$$

где статистика $Z$ уже встречалась нам при изучении мультиномиальной модели. Однако ее распределение в двух моделях различается.

Поскольку случайные величины $a$ и $c$ независимы и имеют биномиальные распределения, то дисперсия статистики $q_1^* - q_2^*$ равна

$$D(q_1^* - q_2^*) = D(q_1^*) + D(q_2^*) = \frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}. \quad (34)$$

При справедливости гипотезы однородности (32) по теореме Муавра-Лапласа статистика

$$\frac{q_1^* - q_2^*}{\sqrt{D(q_1^* - q_2^*)}}$$

является асимптотически нормальной с математическим ожиданием 0 и дисперсией 1. Пусть $D*(q_1^* - q_2^*)$ - оценка дисперсии $D(q_1^* - q_2^*)$ такая, что при безграничном росте объемов обеих выборок

$$\lim_{n_1 \to \infty, n_2 \to \infty} \frac{D*(q_1^* - q_2^*)}{D(q_1^* - q_2^*)} = 1 \quad (35)$$

(сходимость по вероятности). Тогда асимптотическое правило принятия решения при проверке гипотезы однородности (32) на уровне значимости $\alpha$ таково: если

$$\frac{|q_1^* - q_2^*|}{\sqrt{D*(q_1^* - q_2^*)}} \le C(\alpha), \quad (36)$$

то гипотезу однородности (32) принять, в противном случае отклонить. Как и ранее, здесь

$$C(\alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

где $\Phi^{-1}$ - функция, обратная к функции стандартного нормального распределения [1]. Наиболее распространенному при анализе статистических данных значению уровня значимости $\alpha$ = 0,05 соответствует коэффициент $C(0,05) = 1,96$.

Статистику $D*(q_1^* - q_2^*)$ - оценку дисперсии $D(q_1^* - q_2^*)$ - можно сконструировать разными способами.

В учебниках [] оценку дисперсии получают, заменяя в правой части (34) неизвестные вероятности $q_1$ и $q_2$ на их оценки $q_1^*$ и $q_2^*$. В этом случае критерий (36) построен на основе статистики

$$S = \frac{|q_1^* - q_2^*|}{\sqrt{D^*(q_1^* - q_2^*)}} = \frac{|q_1^* - q_2^*|}{\sqrt{\frac{q_1^*(1-q_1^*)}{n_1} + \frac{q_2^*(1-q_2^*)}{n_2}}}. \quad (37)$$

Переходя к величинам в клетках таблицы четырех полей, получаем

$$S = \frac{|ad - bc|}{\sqrt{\frac{ab(c+d)^3 + cd(a+b)^3}{(a+b)(c+d)}}} \quad (38)$$

*Пример 6.* Для данных табл. 2 оценки вероятностей таковы:

$$q_1^* = \frac{a}{n_1} = \frac{300}{450} = 0{,}667, \; q_2^* = \frac{c}{n_2} = \frac{100}{480} = 0{,}208.$$

По формуле (37)

$$S = \frac{|q_1^* - q_2^*|}{\sqrt{\frac{q_1^*(1-q_1^*)}{n_1} + \frac{q_2^*(1-q_2^*)}{n_2}}} = \frac{|0{,}667 - 0{,}208|}{\sqrt{\frac{0{,}667(1-0{,}667)}{450} + \frac{0{,}208(1-0{,}208)}{480}}} =$$

$$= \frac{0{,}459}{\sqrt{0{,}000494 + 0{,}000343}} = \frac{0{,}459}{\sqrt{0{,}000837}} = \frac{0{,}459}{0{,}0289} = 15{,}89$$

Гипотеза однородности отклоняется как на уровне значимости $\alpha = 0{,}05$, так и на любом другом используемом в практических расчетах уровне значимости.

*Пример 7.* Для данных табл.5 оценки вероятностей таковы:

$$q_1^* = \frac{a}{a+b} = \frac{14}{17} = 0{,}824, \; q_2^* = \frac{c}{c+d} = \frac{6}{10} = 0{,}6.$$

По формуле (38)

$$S = \frac{|14 \times 4 - 3 \times 6|}{\sqrt{\frac{14 \times 3 \times 10^3 + 6 \times 4 \times 17^3}{17 \times 10}}} = \frac{38}{\sqrt{\frac{42000 + 117912}{170}}} = \frac{38}{\sqrt{940{,}659}} = \frac{38}{30{,}6702} = 1{,}239$$

Гипотеза однородности принимается.

Приведенные в настоящем разделе правила проверки гипотезы однородности двух биномиальных распределений основаны на

асимптотической нормальности рассматриваемых статистик. Если в клетках таблицы четырех полей стоят сравнительно небольшие числа, то вместо предельных соотношений целесообразно пользоваться результатами для конечных объемов выборок. Например, таблицами и рекомендациями по их использованию в классическом сборнике [10, табл. 5.6].

**12. Выводы**

Таблицу четырех полей строят для описания совместного распределения двух альтернативных (бинарных, дихотомических) признаков. Она является одной из простейших объектов изучения в статистике нечисловых данных [18, 19]. Однако методы статистического анализа данных, собранных в таблице четырех полей, в литературе (включая Интернет-источники) не всегда представлены адекватно, их рассматривают неполно или с ошибками. Настоящая статья посвящена рассмотрению таких методов в соответствии с традициями отечественной школы теории вероятностей и математической статистики, основанной А.Н. Колмогоровым.

Статистический анализ таблиц четырех полей должен начинаться с выбора модели порождения данных. Применяют мультиномиальную модель и модель двух выборок. Проверяемые гипотезы и правила принятия решений меняются при переходе от одной модели к другой.

Нельзя обоснованно судить о наличии связи между признаками только по величине тех или иных коэффициентов. Необходимо применять подходы теории проверки статистических гипотез. В мультиномиальной модели проверяют гипотезу независимости, а в модели двух выборок - гипотезу однородности долей. Только при отклонении нулевой гипотезы можно говорить о наличии связи между признаками, соответственно, о наличии эффекта при переходе от одной выборки к другой.

В настоящей статье разработан метод вычисления асимптотических распределений функций от чисел в клетках таблицы четырех полей. Он состоит в применении многомерной центральной предельной теоремы теории вероятностей и метода линеаризации функций, который на основе состоятельности оценок вероятностей с помощью частот позволяет выделить распределение главного члена как основной составляющей функций от чисел в клетках таблицы четырех полей [1].

Проверка статистических гипотез основана на использовании дисперсий коэффициентов ассоциации, коллигации и контингенции в мультиномиальной модели и разности выборочных долей в модели двух выборок. В применении дисперсий проявляется преимущество нашего подхода по сравнению с распространенной традицией.

В различных материалах, размещенных в Интернете, а также в печатных публикациях имеется некорректная фраза: "Считается, что если коэффициент ассоциации превосходит 0,5 и коэффициент контингенции больше 0,3, то это свидетельствует о существенной связи между признаками". Приведенная фраза некорректна прежде всего потому, что игнорирует возможность получения отрицательных значений коэффициентов ассоциации и контингенции, достаточно больших по абсолютной величине. От этой некорректности легко избавиться, заменив в приведенных выше неравенствах значения коэффициентов на их абсолютные значения. Более существенно, что говорить о "существенной связи между признаками" можно говорить лишь тогда, когда отклонена гипотеза независимости. В случаях, когда в клетках таблицы четырех полей стоят сравнительно небольшие числа, может случиться так, что хотя рассматриваемые коэффициенты заметно отличаются от 0, но тем не менее гипотеза независимости отклоняется.

Рекомендации, полученные в настоящей статье, основаны на асимптотической нормальности рассматриваемых коэффициентов. Если в

клетках таблицы четырех полей стоят сравнительно небольшие числа, то вместо предельных соотношений целесообразно пользоваться результатами для конечных объемов выборок, в частности, таблицами в классическом сборнике "Таблицы математической статистики" Л.Н. Большева и Н.В. Смирнова [10, табл. 5.6] или соответствующими компьютерными программами.

**Литература**

1. Орлов А.И. Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
2. Орлов А.И. Статистика нечисловых данных - центральная часть современной прикладной статистики // Научный журнал КубГАУ. 2020. № 156. С. 111–142.
3. Орлов А.И. Вероятность и прикладная статистика: основные факты: справочник. — М.: КноРус, 2015. — 190 с.
4. Орлов А.И. Вероятностно-статистические модели данных - основа методов прикладной статистики // Заводская лаборатория. Диагностика материалов. 2020. Т.86. № 7. С. 5-6.
5. Уилкс С. Математическая статистика. - М.: Мир, 1967. -632 с.
6. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи / Пер. с англ. Л.И. Гальчука, А.Т. Терехина ; Под ред. А. Н. Колмогорова. - Москва : Наука, 1973. - 899 с.
7. Yule G.U. On the association of attributes in statistics // Philosophical Transactions of the Royal Society of London. Series A. 1900. V. 194. Pp. 257–319.
8. Yule G.U. On the methods of measuring association between two attributes // Journal of the Royal Statistical Society. 1912. V. 75. No. 6. Pp. 579-652.
9. Pearson E.S. The choice of statistical tests illustrated on the interpretation of data classed in a 2x2 table // Biometrika. 1947. V. 34. Pp. 139-167.
10. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука. Главная редакция физико-математическом литературы, 1983. —416 с.
11. Аптон Г. Анализ таблиц сопряженности. - М. : Финансы и статистика, 1982. - 143 с.
12. Орлов А.И. Метод проверки гипотез по совокупности малых выборок и его применение в теории статистического контроля // Научный журнал КубГАУ. 2014. №104. С. 38–52.
13. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 3-е, исправленное и дополненное. – М.: Экзамен, 2004. – 576 с.
14. Орлов А.И. Организационно-экономическое моделирование : учебник : в 3 ч. Ч.3. Статистические методы анализа данных. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2012. – 624 с.
15. Орлов А.И. Многообразие критериев проверки однородности двух независимых выборок / Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. Вып.29. - Пермь: Перм. гос. нац. иссл. ун-т, 2019. - С. 64-83.
16. Орлов А.И. О методах проверки однородности двух независимых выборок // Заводская лаборатория. Диагностика материалов. 2020. Т.86. № 3. С. 67-76.

17. Орлов А.И. Система моделей и методов проверки однородности двух независимых выборок / Научный журнал КубГАУ. 2020. №157. С. 145 – 169.
18. Орлов А.И. Организационно-экономическое моделирование: : учебник : в 3 ч. Ч.1: Нечисловая статистика. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2009. — 542 с.
19. Орлов А.И. Статистика нечисловых данных - центральная часть современной прикладной статистики // Научный журнал КубГАУ. 2020. № 156. С. 111–142.

## References

1. Orlov A.I. Prikladnaya statistika. — M.: Ekzamen, 2006. — 671 s.
2. Orlov A.I. Statistika nechislovyh dannyh - central'naya chast' sovremennoj prikladnoj statistiki // Nauchnyj zhurnal KubGAU. 2020. № 156. S. 111–142.
3. Orlov A.I. Veroyatnost' i prikladnaya statistika: osnovnye fakty: spravochnik. — M.: KnoRus, 2015. — 190 s.
4. Orlov A.I. Veroyatnostno-statisticheskie modeli dannyh - osnova metodov prikladnoj statistiki // Zavodskaya laboratoriya. Diagnostika materialov. 2020. T.86. № 7. S. 5-6.
5. Uilks S. Matematicheskaya statistika. - M.: Mir, 1967. -632 s.
6. Kendall M.Dzh., St'yuart A. Statisticheskie vyvody i svyazi / Per. s angl. L.I. Gal'chuka, A.T. Terekhina ; Pod red. A. N. Kolmogorova. - Moskva : Nauka, 1973. - 899 s.
7. Yule G.U. On the association of attributes in statistics // Philosophical Transactions of the Royal Society of London. Series A. 1900. V. 194. Pp. 257–319.
8. Yule G.U. On the methods of measuring association between two attributes // Journal of the Royal Statistical Society. 1912. V. 75. No. 6. Pp. 579-652.
9. Pearson E.S. The choice of statistical tests illustrated on the interpretation of data classed in a 2x2 table // Biometrika. 1947. V. 34. Rp. 139-167.
10. Bol'shev L.N., Smirnov N.V. Tablicy matematicheskoj statistiki. - M.: Nauka. Glavnaya redakciya fiziko-matematicheskom literatury, 1983. —416 s.
11. Apton G. Analiz tablic sopryazhennosti. - M. : Finansy i statistika, 1982. - 143 c.
12. Orlov A.I. Metod proverki gipotez po sovokupnosti malyh vyborok i ego primenenie v teorii statisticheskogo kontrolya // Nauchnyj zhurnal KubGAU. 2014. №104. S. 38–52.
13. Orlov A.I. Ekonometrika. Uchebnik dlya vuzov. Izd. 3-e, ispravlennoe i dopolnennoe. – M.: Ekzamen, 2004. – 576 s.
14. Orlov A.I. Organizacionno-ekonomicheskoe modelirovanie : uchebnik : v 3 ch. CH.3. Statisticheskie metody analiza dannyh. – M.: Izd-vo MGTU im. N.E. Baumana, 2012. – 624 s.
15. Orlov A.I. Mnogoobrazie kriteriev proverki odnorodnosti dvuh nezavisimyh vyborok / Statisticheskie metody ocenivaniya i proverki gipotez: mezhvuz. sb. nauch. tr. Vyp.29. - Perm': Perm. gos. nac. issl. un-t, 2019. - S. 64-83.
16. Orlov A.I. O metodah proverki odnorodnosti dvuh nezavisimyh vyborok // Zavodskaya laboratoriya. Diagnostika materialov. 2020. T.86. № 3. S. 67-76.
17. Orlov A.I. Sistema modelej i metodov proverki odnorodnosti dvuh nezavisimyh vyborok / Nauchnyj zhurnal KubGAU. 2020. №157. S. 145 – 169.
18. Orlov A.I. Organizacionno-ekonomicheskoe modelirovanie: : uchebnik : v 3 ch. CH.1: Nechislovaya statistika. — M.: Izd-vo MGTU im. N. E. Baumana, 2009. — 542 s.
19. Orlov A.I. Statistika nechislovyh dannyh - central'naya chast' sovremennoj prikladnoj statistiki // Nauchnyj zhurnal KubGAU. 2020. № 156. S. 111–142.