

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики

На правах рукописи

ЗУБОВ Василий Николаевич

**О СКОРОСТИ СХОДИМОСТИ СТАТИСТИК КРИТЕРИЕВ СОГЛАСИЯ СО
СТЕПЕННЫМИ МЕРАМИ РАСХОЖДЕНИЯ К ХИ-КВАДРАТ РАСПРЕДЕЛЕНИЮ**

Специальность 01.01.05 - "Теория вероятностей и математическая
статистика"

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2010

Работа выполнена на кафедре математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова.

Научный руководитель: доктор физико-математических наук,
профессор Ульянов Владимир Васильевич

Официальные оппоненты: доктор физико-математических наук,
профессор Богачев Владимир Игоревич

кандидат физико-математических наук,
Хохлов Владимир Иванович

Ведущая организация: Московский государственный институт
электроники и математики

Защита диссертации состоится 29 октября 2010 года в 11 часов на заседании диссертационного совета Д 501.001.44 в Московском государственном университете имени М. В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет ВМиК, аудитория 685.

С диссертацией можно ознакомиться в библиотеке факультета ВМиК МГУ. С текстом автореферата можно ознакомиться на официальном сайте факультета ВМиК Московского государственного университета имени М.В. Ломоносова

<http://www.cmc.msu.ru> в разделе „Наука “ – „Работа диссертационных советов“ – „Д 501.001.44“.

Автореферат разослан __ сентября 2010 года.

Ученый секретарь
диссертационного совета
профессор

Н. П. Трифонов

Общая характеристика работы

Актуальность темы

На протяжении многих десятилетий критерии согласия, предназначенные для проверки соответствия имеющихся выборочных данных заданному распределению, не теряют своей значимости в математической статистике. Выдающиеся результаты, полученные в этом направлении, связаны с именами А. Н. Колмогорова, Н. В. Смирнова, К. Пирсона. Один из подходов состоит в следующем: данные выборки группируются на основе попадания в непересекающиеся подмножества (ячейки) области возможных значений элементов выборки, с тем чтобы затем сопоставить частоты попадания в эти подмножества с теоретическими вероятностями, которые могут быть вычислены. При этом проверка соответствия выборки исходному распределению заменяется проверкой соответствия сгруппированных данных полиномиальному распределению. Несомненное достоинство такого подхода состоит в его универсальности.

Классический критерий согласия, предложенный К. Пирсоном и основанный на вышеупомянутой методологии, использует так называемую статистику χ^2 . Эта статистика имеет простой вид и удобна в применении. Вместе с тем, для получения хорошей точности с помощью этого критерия необходимо иметь достаточно большой объем входных данных (в сумме и по отдельности в каждой ячейке). Кроме того, на практике чаще всего приходится заменять распределение статистики асимптотическим. Аппроксимирующее распределение зависит от числа ячеек, а величина ошибки чаще всего неизвестна. Непонятно и то, является ли статистика χ^2 оптимальной на малых объемах выборки.

В связи с этим многие ученые исследовали другие подходы к построению критериев согласия с целью найти наиболее эффективный в том или ином статистическом смысле. Здесь можно упомянуть работы С. Фейнберга, Х. Ланкастера, Д. Мура, Г. Вотсона. Неплохой сводный анализ различных альтернатив приведен в сводной работе С. Хорна¹. Особое место в этих исследованиях принадлежит работам Н. Крисси² и Т. Риды³. Эти авторы ввели в употребление и произвели первичный анализ семейства *степенных статистик согласия*, предназначенного для построения критериев согласия по сгруппированным данным с использованием степенных мер расхождения между эмпирическими частотами и теоретическими вероятностями. Семейство параметризо-

¹S. HORN. Goodness-of-Fit Tests for Discrete Data: A Review and an Application to a Health Impairment Scale. *Biometrics*, **33**, 1, pp. 237–247, 1977.

²N. A. C. CRESSIE, T. R. C. READ. Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B*, **46**, No. 3, 440–464, 1984.

³T. R. C. READ. Small sample comparisons for the power divergence goodness-of-fit statistics, *Journal of the American Statistical Association*, **79**, 388, pp. 929–935, 1984.

вано вещественным параметром λ , при этом как собственно статистика χ^2 , так и часто используемые статистики являются частными случаями.

Хотя при фиксированной процедуре группировки и количестве ячеек все статистики семейства асимптотически эквивалентны (имеют одинаковое предельное хи-квадрат распределение), на основании имеющихся работ⁴ можно утверждать, что в определенных ситуациях классическая статистика хи-квадрат не является наиболее эффективной в рамках семейства. Для практической реализации альтернативных критериев (например, при вычислении критических значений) необходимо четко понимать, насколько хорошо статистики семейства аппроксимируются предельным хи-квадрат распределением и как это соотносится с точностью аппроксимации для статистики χ^2 . В связи с этим актуальной является задача исследования степенных статистик согласия на предмет скорости их [слабой] сходимости к хи-квадрат распределению (при выполнении основной гипотезы и фиксированном количестве ячеек группировки).

Из работ Ф. Гетце, В. Ульянова и Дж. Ярнольда известны оценки скорости сходимости в частном случае статистики χ^2 ; также в работах Т. Риды⁵, М. Шиотани и Я. Фуджикоши⁶ получены различные асимптотические разложения функции распределения степенных статистик согласия. Тем не менее, ни одна из этих работ не позволяет построить оценки скорости сходимости последних к хи-квадрат распределению.

Цель работы

Целью настоящей работы является получение оценок скорости слабой сходимости степенных статистик согласия к хи-квадрат распределению.

Научная новизна

Все полученные в диссертации результаты являются новыми и состоят в следующем:

1. для всех степенных статистик согласия впервые получены степенные оценки скорости слабой сходимости к хи-квадрат распределению;

⁴N. A. C. CRESSIE, T. R. C. READ. Multinomial goodness-of-fit tests, Journal of the Royal Statistical Society, Series B, **46**, No. 3, 440–464, 1984.

T. R. C. READ. Small sample comparisons for the power divergence goodness-of-fit statistics, Journal of the American Statistical Association, **79**, 388, pp. 929–935, 1984.

⁵T. R. C. READ. Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics., The Annals of Mathematical Statistics, **36**, Part A, 59–69, 1984.

⁶M. SIOTANI AND Y. FUJIKOSHI. Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics, Hiroshima Math. J., **14** 1984 г., 115–124; technical report of the Hiroshima statistical research group, 1980.

2. для случая трех ячеек группировки получены более точные оценки, которые превосходят ранее известные оценки для статистики хи-квадрат;
3. в процессе работы разработан новый метод построения оценок скорости сходимости степенных статистик согласия, основанный на сведении исходной задачи к задаче о приближении числа точек решетки в заданном выпуклом множестве его нормированным объемом.

Методы исследования

Основной метод состоит в последовательном сведении задачи из математической статистики к задаче из теории чисел о приближении числа целых точек в некотором выпуклом множестве его нормированным объемом. В последующем привлекаются имеющиеся результаты из теории чисел, при этом применимость этих результатов к проблеме диссертации обосновывается с помощью методов теории оптимального управления, тензорного исчисления, и дифференциальной геометрии.

Теоретическая и практическая значимость

Результаты работы имеют теоретический характер. Они представляют значительный интерес при изучении асимптотических свойств степенных статистик согласия. Практическая значимость исследования связана с поиском эффективных альтернатив традиционным критериям согласия.

Апробация работы и публикации

Автор имеет 6 публикаций по тематике диссертации. Работа докладывалась на научно-исследовательском семинаре кафедры математической статистики факультета ВМиК МГУ (март 2009 г.), на Российско-японском симпозиуме [Математического Института им В. А. Стеклова РАН] "Стохастический анализ сложных статистических моделей"(сентябрь 2009 г.), на Российско-японском семинаре [Математического Института им В. А. Стеклова РАН] "Сложные стохастические модели: асимптотики и приложения"(июнь 2007 г.), а также на 28-й Конференции Европейских Статистиков (август 2010 г., Университет Пирея, Греция). Основные результаты были опубликованы в журналах «Теория вероятностей и ее применения», «Обзор прикладной и промышленной математики» и в Hiroshima Mathematical Journal.

Структура диссертации

Работа общим объемом 98 страниц состоит из введения, двух глав, заключения и приложения. Во введении в более технических терминах обсуждается постановка задачи, проводится краткий анализ имеющихся результатов, формулируются основные идеи и методы доказательства. В главе 1 рассматривается частный случай, когда число ячеек группировки равно трем. В главе 2 последнее ограничение снимается. Наконец в приложение выведены некоторые вспомогательные результаты технического характера. Список литературы состоит из 54 наименований.

Краткое содержание диссертации

Вводные замечания

Мы рассматриваем вектор $\mathbf{Y} = (Y_1, \dots, Y_k)^T$, имеющий полиномиальное распределение $M_k(n, \pi)$, т. е.

$$\Pr(Y_1 = n_1, \dots, Y_k = n_k) = \begin{cases} n! \prod_{j=1}^k (\pi_j^{n_j} / n_j!), & n_j = 0, 1, \dots, n \ (j = 1, \dots, k) \\ & \text{и } \sum_{j=1}^k n_j = n, \\ 0 & \text{иначе} \end{cases}$$

где $\pi = (\pi_1, \dots, \pi_k)^T$, $\pi_j > 0$, $\sum_{j=1}^k \pi_j = 1$. Мы предполагаем, что число ячеек группировки k фиксировано. Далее будем считать выполненной основную гипотезу $H_0: \pi = \mathbf{p}$. Если изначальное распределение было отлично от полиномиального, вектор $\mathbf{p}$ состоит из теоретических вероятностей попадания случайной величины, имеющей исходное распределение, в соответствующие ячейки.

Основным объектом изучения в настоящей работе является построенное по вышеуказанному распределению семейство *степенных статистик согласия*

$$t_\lambda(\mathbf{Y}) = \frac{2}{\lambda(\lambda + 1)} \sum_{j=1}^k Y_j \left[\left(\frac{Y_j}{np_j} \right)^\lambda - 1 \right], \quad \lambda \in \mathbb{R}, \quad (1)$$

З а м е ч а н и е 1. При $\lambda = 0, -1$ эту запись следует понимать как результат предельного перехода.

Предполагая выполненной основную гипотезу, рассмотрим преобразование

$$X_j = (Y_j - np_j) / \sqrt{n}, \quad j = 1, \dots, k, \quad r = k - 1, \quad \mathbf{X} = (X_1, \dots, X_r)^T. \quad (2)$$

Компоненты вектора $\mathbf{X}$ сосредоточены на решетке вида

$$L = \{\mathbf{x} = (x_1, \dots, x_r)^T; \mathbf{x} = (\frac{1}{\sqrt{n}})(\mathbf{m} - n\mathbf{p}), \mathbf{p} = (p_1, \dots, p_r)^T, \mathbf{m} = (n_1, \dots, n_r)^T\},$$

где n_j — неотрицательные целые числа. При этом, поскольку компоненты вектора $\mathbf{Y}$ в сумме всегда равны n , мы можем положить $X_k = -(X_1 + \dots + X_r)$. Величина r , таким образом, определяет размерность нашей задачи.

Для того чтобы исследовать статистики $t_\lambda(\mathbf{Y})$ на слабую сходимость, мы рассматриваем их как функции от $\mathbf{X}$ (в дальнейшем обозначаются как $T_\lambda(\mathbf{X})$). Заметим, что функцию распределения $\mathbf{P}(T_\lambda(\mathbf{X}) < c)$ можно записать в виде вероятности попадания случайного вектора $\mathbf{X}$, имеющего решетчатое распределение, в некоторое множество $B^\lambda(c)$ (в дальнейшем будет обозначаться как B^λ , проекции на одну из координат будут обозначаться дополнительным нижним индексом). Одним из важных источников по оценке такого рода вероятностей является работа Дж. Ярнольда⁷. Его результаты, доказанные с привлечением многомерных разложений Эджворта, можно применить к оценке искомой вероятности. При этом мы получим

$$\Pr(\mathbf{X} \in B^\lambda) = J_1 + J_2 + O(n^{-1}). \quad (3)$$

Члены этого разложения имеют следующий вид

$$J_1 = \int \dots \int_{B^\lambda} \phi(\mathbf{x}) \left\{ 1 + \frac{1}{\sqrt{n}} h_1(\mathbf{x}) + \frac{1}{n} h_2(\mathbf{x}) \right\} d\mathbf{x}, \text{ где} \quad (4)$$

$$h_1(\mathbf{x}) = -\frac{1}{2} \sum_{j=1}^k \frac{x_j}{p_j} + \frac{1}{6} \sum_{j=1}^k x_j \left(\frac{x_j}{p_j} \right)^2, \quad (5)$$

$$h_2(\mathbf{x}) = \frac{1}{2} h_1(\mathbf{x})^2 + \frac{1}{12} \left(1 - \sum_{j=1}^k \frac{1}{p_j} \right) + \frac{1}{4} \sum_{j=1}^k \left(\frac{x_j}{p_j} \right)^2 - \frac{1}{12} \sum_{j=1}^k x_j \left(\frac{x_j}{p_j} \right)^3; \quad (6)$$

$$J_2 = -\frac{1}{\sqrt{n}} \sum_{l=1}^r n^{-\frac{(r-l)}{2}} \sum_{x_{l+1} \in L_{l+1}} \dots \sum_{x_r \in L_r} \left[\int \dots \int_{B_l^\lambda} [S_1(\sqrt{n}x_l + np_l)\phi(\mathbf{x})]_{\lambda_l(x^*)}^{\theta_l(x^*)} dx_1, \dots, dx_{l-1} \right]; \quad (7)$$

$$L_j = \{\mathbf{x}: x_j = (\frac{1}{\sqrt{n}}(n_j - np_j)), \text{ } n_j \text{ и } p_j \text{ определены как и раньше}\}; \quad (8)$$

$$S_1(x) = x - [x] - \frac{1}{2}, \quad [x] \text{ — целая часть } x; \quad (9)$$

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{r}{2}} |\Omega|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} \mathbf{x}^T \Omega^{-1} \mathbf{x} \right). \quad (10)$$

⁷J. K. YARNOLD. Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set, The Annals of Mathematical Statistics, **43**, No. 5, 1566–1580, 1972.

В последней формуле матрица $\Omega = (\delta_i^j p_i - p_i p_j) \in \mathbb{R}^{(k-1) \times (k-1)}$ есть ковариационная матрица вектора $\mathbf{X}$. В формуле (7) функции $\lambda_l(x^*)$ и $\theta_l(x^*)$ обозначают диапазон изменения координаты x_l внутри множества B^λ , если зафиксировать остальные координаты. Величину J_1 можно интуитивно понимать как разложение Эджворта до порядка $O(\frac{1}{n})$ для непрерывной части распределения, в то время как величина J_2 появляется как дополнительный корректирующий член для решетчатых распределений. Именно оценка этого члена представляет наибольшую трудность с практической точки зрения.

Вначале были получены результаты для классического случая $\lambda = 1$ (статистика хи-квадрат). Нетрудно показать, что множество B^1 в этом простом случае представляет собой эллипсоид. В той же работе [7] Дж. Ярнхольд сумел упростить сумму слагаемых J_1 и J_2 в этом простом случае и привел разложение (3) к виду

$$\Pr(\mathbf{X} \in B^1) = K_r(c) + (N^1 - n^{\frac{r}{2}} V^1) e^{-\frac{c}{2}} / ((2\pi n)^r \prod_{j=1}^k p_j)^{\frac{1}{2}} + O(n^{-1}), \quad (11)$$

где $K_r(c)$ — функция распределения хи-квадрат с r степенями свободы, N^1 — число точек решетки L в множестве B^1 , V^1 — объем множества B^1 . Опираясь на результат Эссеена для эллипсоидов, он получил оценку для второго слагаемого (11) вида $O(n^{-\frac{r}{r+1}})$.

Позднее М. Шиотани, Я. Фуджикоши [6] и Т. Рид [5] получили аналог этих результатов для произвольного λ , а именно, что

$$\begin{aligned} J_1 &= K_r(c) + O(n^{-1}) \\ J_2 &= (N^\lambda - n^{\frac{r}{2}} V^\lambda) e^{-\frac{c}{2}} / ((2\pi n)^r \prod_{j=1}^k p_j)^{\frac{1}{2}} + o(1), \\ V^\lambda &= V^1 + O\left(\frac{1}{n}\right). \end{aligned} \quad (12)$$

Заметим, что в отличие от работы Дж. Ярнхольда остаточный член в (12) не оценен. Это означает, что для того чтобы оценить скорость сходимости к хи-квадрат распределению автору нужно было, во-первых, получить содержательную оценку остаточного члена в (12) и, во-вторых, оценить главную часть J_2 .

Проблема оценки главной части J_2 сводится к проблеме оценки разности числа точек решетки, попадающих в множество B^λ , и нормированного объема этого множества. С помощью преобразования масштаба задача сводится к оценке разности числа точек с целочисленными координатами, попадающих в множество, полученное линейным расширением из B^λ , и объема этого расширенного множества. Такая задача является ничем иным, как обобщением известной из теории чисел задачи Гаусса. Для наших целей (см. формулу (8)) при переходе

от целых точек к точкам решетки необходимо взять нормировочный коэффициент $\sqrt{n}$.

Общий алгоритм решения состоит в следующем. Вначале мы оцениваем остаточный член в (12), используя близость множества $B^\lambda = \{\mathbf{x} \mid T_\lambda(\mathbf{x}) < c\}$ и эллипсоида B^1 , соответствующего статистике хи-квадрат. Это означает, что, отталкиваясь от формулы (7), мы можем заменить функции $\theta_l(x^*)$ и $\lambda_l(x^*)$ на функции $\hat{\theta}_l(x^*)$ и $\hat{\lambda}_l(x^*)$, соответствующие границе множества B^1 . Поскольку $\phi(\mathbf{x})$ на границе B^1 есть константа, мы можем упростить подынтегральное выражение в (7) с некоторой погрешностью, которую можно подсчитать. Оказывается, что величина, остающаяся ведущим членом после этих преобразований, настолько похожа на соответствующий член из разложения для эллипсоидов, что к ней применимы рассуждения работы Ярнольда [7] (стр. 1571-1572), позволяющие свести оценку J_2 к оценке величины

$$(N^\lambda - n^{\frac{r}{2}} V^\lambda) e^{-\frac{c}{2}} / ((2\pi n)^r \prod_{j=1}^k p_j)^{\frac{1}{2}}.$$

На втором этапе мы рассматриваем линейное расширение множества B^λ с фактором $\sqrt{n}$, соответствующим преобразованию координат (2). Мы проверяем условия применимости соответствующей теоремы из теории чисел и получаем искомый результат.

Глава I

В этой главе мы исследуем проблему для случая, когда число ячеек группировки k равно трем. Это исследование предшествовало исследованию статистик произвольной размерности и было предназначено для выработки методологии решения задачи. Тем не менее полученные здесь результаты заслуживают отдельного внимания не только в силу того, что не дублируются в полной мере во второй части, но и в силу своеобразных методов решения, существенно отличных от методов решения общей проблемы.

Конкретизируя рассуждения, изложенные выше (см. формулы (1)–(10)), имеем

$$B^\lambda = \{(x, y) : T_\lambda(x, y) < c\},$$

и

$$\begin{aligned} T_\lambda(x, y) = & \frac{2}{\lambda(\lambda+1)}(np_1 + \sqrt{n}x) \left[\left(1 + \frac{x}{\sqrt{n}p_1}\right)^\lambda - 1 \right] \\ & + \frac{2}{\lambda(\lambda+1)}(np_2 + \sqrt{n}y) \left[\left(1 + \frac{y}{\sqrt{n}p_2}\right)^\lambda - 1 \right] \\ & + \frac{2}{\lambda(\lambda+1)}(np_3 - \sqrt{n}(x+y)) \left[\left(1 - \frac{x+y}{\sqrt{n}p_3}\right)^\lambda - 1 \right]. \end{aligned}$$

Основные результаты этой главы были опубликованы в работах [1] и [3]. Используя упрощенное разложение (12), полученное Шиотани, Фуджикоши и Ридом, была доказана следующая основная теорема (в автореферате сохранена оригинальная нумерация диссертации).

Теорема 3 *Для любого $\lambda \in \mathbb{R}$ справедлива оценка*

$$J_2(B^\lambda) = O\left(n^{-100/146}(\log n)^{315/146}\right). \quad (13)$$

Как и раньше, мы будем оперировать с сечениями множеств по одной из координат, которые будут обозначаться дополнительным нижним индексом. Пусть $\hat{\theta}_1$ и $\hat{\lambda}_1$ — функции, обозначающие диапазон изменения переменной x внутри B^1 при фиксированном y , B^1_1 — область определения этих функций. Напомним, что функции θ_1 и λ_1 соответствуют множествам B^λ . Также определим

$$T_1(x, y) = \left(\frac{1}{p_1} + \frac{1}{p_3}\right)x^2 + \frac{2}{p_3}xy + \left(\frac{1}{p_2} + \frac{1}{p_3}\right)y^2.$$

Согласно алгоритму вначале в разделе 1.2 диссертации мы редуцируем член J_2 к упрощенному виду, проводя необходимую оценку погрешности. Для этого необходимы следующие оценки

Лемма 1 *Мера Лебега множества $B^\lambda_1 \setminus B^1_1$ имеет порядок $O\left(n^{-\frac{1}{2}}\right)$.*

Лемма 2 *Существуют константы $c_1 > 0, c_2 > 0$ такие, что*

$$\begin{aligned} \left|\theta_1(y) - \hat{\theta}_1(y)\right| &\leq c_1 n^{-\frac{1}{4}}, \\ \left|\lambda_1(y) - \hat{\lambda}_1(y)\right| &\leq c_2 n^{-\frac{1}{4}} \end{aligned}$$

для всех $y \in B^\lambda_1 \cap B^1_{1-}$, начиная с $N = \lceil (cp_2(p_1 + p_3))^{-1} \rceil$.

Эти две леммы были получены Ж. Асылбековым в работе [3]. Дальнейшие рассуждения позволяют показать, что справедлива

Теорема 4 *Член $J_2(B^\lambda)$, определяемый формулой (7), можно представить в следующем виде*

$$J_2(B^\lambda) = \frac{d}{n}(N^\lambda - nV^\lambda) + O(n^{-\frac{3}{4}}), \quad (14)$$

где $d > 0$ — константа.

Далее в разделе 1.3 для оценки (14) применяется следующий результат, полученный М. Хаксли⁸:

Теорема 1 (Хаксли, 1993) *Пусть B — выпуклая евклидова плоская область площади A , ограниченная простой замкнутой кривой C , состоящей из конечного числа частей C_i , каждая из которых*

⁸М. N. HUXLEY. Exponential sums and lattice points II, Proceedings of London Mathematical Society, **3**, 66, 279–301, 1993.

три раза непрерывно-дифференцируема в следующем смысле: радиус кривизны ρ непрерывен относительно угла смежности (тангенциального угла) ψ и не равен нулю на каждой части C_i , а также непрерывно-дифференцируем относительно угла смежности. Пусть число M достаточно велико, и пусть MV обозначает множество, образованное увеличением множества V линейно в M раз. Тогда для любого изометрического вложения множества MV в евклидову плоскость число целых точек (m, n) в MV есть

$$AM^2 + O\left(IM^{46/73}(\log M)^{315/146}\right),$$

где I – число, зависящее от кривой C , но не от M и не от вложения множества MV .

Интерпретация этой теоремы, равно как и других утверждений такого рода, достаточно проста. Если выпуклое множество обладает достаточно гладкой регулярной границей, то количество целых точек, попадающих в его линейное расширение, может быть с хорошей степенью точности приближено площадью этого линейного расширения. Дополнительно теорема утверждает следующее.

Теорема 1 (Хаксли, 1993, дополнение) Если помимо вышеуказанного части C_i четырежды непрерывно дифференцируемы в том смысле, что ρ дважды непрерывно дифференцируем по отношению к углу смежности (тангенциальному углу) ψ , то тогда мы можем взять

$$\begin{aligned} I = & \sum_i \min_{C_i} \left(1 + \frac{1}{\rho^2} \left(\frac{d\rho}{d\psi} \right)^2 \right)^{-69/146} \rho^{46/73} \\ & + \sum_i \int_{C_i} \left(1 + \frac{|\rho d^2\rho/d\psi^2|}{\rho^2 + (d\rho/d\psi)^2} \right) \\ & \times \left(1 + \frac{1}{\rho^2} \left(\frac{d\rho}{d\psi} \right)^2 \right)^{-69/146} \left| \frac{d\rho}{d\psi} \right| \rho^{-33/73} d\psi, \end{aligned}$$

при условии, что M достаточно велико для выполнения неравенств

$$M \geq \frac{1}{\rho} \quad \text{и} \quad \frac{1}{\rho^{64}} \left| \frac{d\rho}{d\psi} \right|^{53} \leq M^{11} (\log M)^{387/8}$$

по отдельности на каждом участке C_i .

Это дополнение к теореме существенно для последующих рассуждений, поскольку оно позволяет нам оценить константу I . Особая сложность рассматриваемой в диссертации задачи состоит в том, что в действительности множества B^λ и константа I зависят от n . В связи с этим получение итогового порядка ошибки по n возможно только после получения оценки сверху на I .

Как видно из условия теоремы для ее применимости достаточно показать наличие у границы множества гладкой параметризации. Для этого сначала доказывается строгая выпуклость множеств B^λ (получено Ж. Асылбековым). Затем для функции

$$U(r, t) = T_\lambda(r \cos t, r \sin t) - c,$$

доказывается

Лемма 5

$$\exists s, N: \forall (r, t) \in \partial B^\lambda, n \geq N \quad \frac{\partial U(r, t)}{\partial r} \geq s > 0.$$

Следовательно по теореме о неявной функции справедлива

Лемма 6 Пусть (r_0, t_0) — точка области определения, в которой функция $U(r, t)$ обращается в 0. Тогда для любого достаточно малого положительного числа ε найдется такая окрестность точки t_0 , что в пределах этой окрестности существует единственная функция $r = r(t)$, которая удовлетворяет условию $|r - r_0| < \varepsilon$ и является решением уравнения

$$U(r, t) = 0,$$

причем эта функция $r = r(t)$ непрерывна и пять раз дифференцируема в указанной окрестности точки t_0 .

и, наконец,

Лемма 7 Для кривой

$$T_\lambda(x, y) = c$$

существует четырежды непрерывно-дифференцируемая параметризация в виде $x = x(t) = r(t) \cos t$, $y = y(t) = r(t) \sin t$, $t \in [0, 2\pi]$.

Леммы 6 и 7 получены Ж. Асылбековым. Следующие две леммы позволяют доказать равномерную отделенность полярного радиуса от нуля и оценить константу I из условия теоремы Хаксли.

Лемма 8 Пусть $r_n(t)$ — полярный радиус множества B^λ , а $r(t)$ — полярный радиус множества B^1 . Тогда справедливо неравенство

$$|r_n(t) - r(t)| \leq \frac{C}{\sqrt{n}}$$

Лемма 9 Начиная с некоторого номера N радиус кривизны ρ границы ∂B^λ равномерно (по n) ограничен сверху и равномерно отделен от 0; его первая и вторая производные по тангенциальному углу ψ непрерывны и равномерно ограничены сверху.

Напомним, что N^λ — число точек решетки L , попадающих в множество B^λ . Поскольку решетка L имеет шаг, равный $\frac{1}{\sqrt{n}}$, мы можем рассматривать N^λ как число целых точек в множестве $\sqrt{n}B^\lambda$, которое является линейным увеличением множества B^λ в $\sqrt{n}$ раз. В силу

лемм 4-9 мы можем применить теорему Хаксли к множеству B^λ с линейным фактором $\sqrt{n}$.

Ограниченность константы I следует из оценки сверху

$$I(n) \leq \min_C \rho^{\frac{46}{73}} + \int_C \frac{1 + \left| \frac{d^2 \rho}{d\psi^2} / \rho \right|}{\rho^{\frac{33}{73}}} \left| \frac{d\rho}{d\psi} \right| d\psi$$

и леммы 9. Следовательно, мы можем не учитывать ее при подсчете порядка погрешности. Тогда из теоремы 1 следует, что

$$N^\lambda - nV^\lambda = O\left(n^{46/146}(\log n)^{315/146}\right). \quad (15)$$

Остается подставить (15) в (14), и мы приходим к основной оценке главы I (13).

Глава II

В главе II исследуется проблематика общего случая, когда нет никаких ограничений на количество ячеек группировки. Основные результаты этой главы были опубликованы в работах [2], [4], и [5].

Получена

Теорема 6 *Для члена J_2 справедлива оценка*

$$J_2 = O\left(n^{-1+\frac{1}{r+1}}\right), \quad r \geq 3, \quad (16)$$

Следствие *Для статистики $T_\lambda(\mathbf{X})$ справедливо*

$$\Pr(T_\lambda(\mathbf{X}) < c) = K_r(c) + O\left(n^{-1+\frac{1}{r+1}}\right), \quad r \geq 3.$$

В связи с практической реализацией изложенного во введении алгоритма разбиения задачи на редукцию члена J_2 к упрощенному виду и последующее применение результата из теории чисел необходимо сделать два основных замечания.

Во-первых, в ситуации, изложенной в главе I, при решении первой подзадачи оказывается возможным напрямую использовать близость функций $\theta_l(x^*)$ и $\hat{\theta}_l(x^*)$ ($\lambda_l(x^*)$ и $\hat{\lambda}_l(x^*)$ соответственно). В многомерном варианте из-за наличия в J_2 комбинированных подчленов (см. (7)), содержащих одновременно и суммы, и интегралы, оценка погрешности, полученная этим относительно простым способом, оказывается недостаточно хорошей. Поэтому автору пришлось использовать другой метод, основанный на сглаживающем эффекте гауссовской плотности $\phi(\mathbf{x})$, уменьшающем погрешность аппроксимации.

Во-вторых, результат Хаксли из главы I не имеет прямого обобщения на случай множеств произвольной размерности, поскольку используемая им техника существенно использует тот факт, что мы находимся на плоскости. Поэтому при последующем исследовании многомерного случая был выбран другой результат из теории чисел. В виду больших технических сложностей, связанных с необходимостью проверять все условия применимости результата из теории чисел, автор был вынужден использовать результат, который не является самым точным из имеющихся, но такой, что сложность проверки условий теоремы представлялась приемлемой (это означает, что остается место для уточнения полученных в настоящей работе оценок на основе изложенной методологии).

Именно, в работе используется предложение 9 работы Э. Главки⁹.
Теорема 2 (Э. Главка, 1950) Пусть D — компактное выпуклое множество в $\mathbb{R}^m$, имеющее начало координат своей внутренней точкой. Объем этого множества обозначим через A . Предположим, что границей множества является $m - 1$ -мерная поверхность класса $\mathbb{C}^\infty$, причем всюду на ней гауссова кривизна не равна ни бесконечности, ни нулю. Также предполагается, что определенное специальным образом «каноническое» отображение единичной сферы на D взаимно-однозначно и принадлежит классу гладкости $\mathbb{C}^\infty$. Тогда во множестве, полученном из исходного параллельным переносом на произвольный вектор и линейным расширением в M раз, количество целых точек равно

$$N = AM^m + O\left(IM^{m-2+\frac{2}{m+1}}\right)$$

где величина I зависит только от свойств кривой C , но не от параметров M или A .

Как можно видеть, положив $m = 2$, этот результат дает оценку погрешности $O(M^{\frac{2}{3}})$, в то время как результат Хаксли $O(M^{\frac{46}{73}+\epsilon})$, что лучше. Тем не менее, проверка условий теоремы Главки сложнее, чем проверка условий теоремы Хаксли. Кроме того, в теореме Хаксли константа I не задана в явном виде.

Глава II начинается кратким введением. В разделе 2.2 в общем виде излагается процедура редукции. Здесь в отличие от главы I привлекается ряд терминов и утверждений из дифференциальной геометрии, для которых даны краткие пояснения. Также вводится понятие *максимального сечения* множества.

Определение 1. Назовем сечение множества B^λ *максимальным* по переменной x_1 (*максимальным* в направлении $\mathbf{e}$), если результат ортогонального проектирования этого сечения на плоскость $x_1 = \text{const}$

⁹Е. HLAUKA. Über integrale auf konvexen körpern I. Mh Math **54**, 1–36, 1950.

(на плоскость, ортогональную $\mathbf{e}$), рассматриваемый как множество в $r - 1$ -мерном пространстве, совпадает с проекцией на ту же плоскость всего множества. При этом, очевидно, проекция максимального сечения множества B^λ совпадает с множеством B_1^λ , а множества B^1 — с множеством B_1^1 .

Аппарат максимальных сечений нужен для того, чтобы научиться находить проекции множеств B^λ на координатные оси, необходимые для работы с членом J_2 . Развивая этот аппарат, мы показываем, что справедлива

Лемма 14 Пусть $S = \{\mathbf{x} \mid T(\mathbf{x}) = c\}$ — гладкое $n - 1$ мерное многообразие в $\mathbb{R}^n$ и $\mathbf{e}$ — некоторое направление. Тогда максимальное сечение в направлении $\mathbf{e}$ можно искать из необходимого условия

$$\frac{\partial T(\mathbf{x})}{\partial \mathbf{e}} = 0.$$

Если последнее выполнено, то достаточным условием существования (не обязательно единственного) максимального сечения является одновременное выполнение в каждой точке P на границе сечения равенств

1. $T(\mathbf{x}) = c$,
2. $\partial^2 T(\mathbf{x}) / \partial \mathbf{e}^2 > 0$,
3. минимум $T(\mathbf{x})$ по прямой $\mathbf{x} = P + \mathbf{e}t$ является глобальным по t .

Мы можем работать с множествами B^λ , поскольку справедлива

Лемма 15 В пространстве $\mathbb{R}^r$ множество

$$T_\lambda(\mathbf{x}) = c$$

является $(r - 1)$ -мерным многообразием (поверхностью) класса C^∞ .

Без ограничения общности положим, что мы проектируем множество B^λ на гиперплоскость $x_1 = 0$. Тогда из необходимого условия $\frac{\partial T_\lambda}{\partial x_1} = 0$ и достаточных условий получаем, что независимо от λ максимальное сечение дается уравнением

$$\frac{x_1}{p_1} + \frac{x_1 + \dots + x_r}{p_{r+1}} = 0,$$

Поэтому после несложных вычислений мы получаем, что проекция задается уравнением

$$\frac{c\lambda(\lambda + 1)}{2n} = -1 + \sum_{i=2}^r p_i \left(1 + \frac{x_i}{\sqrt{n}p_i}\right)^{\lambda+1} + (p_1 + p_{r+1}) \left(1 - \frac{x_2 + \dots + x_r}{\sqrt{n}(p_1 + p_{r+1})}\right)^{\lambda+1} \quad (17)$$

Неожиданным и весьма полезным фактом, вытекающим из предыдущих рассуждений является то, что последнее уравнение можно представить в виде $T_\lambda(x') = c$, где $x' = (x_2, \dots, x_r)$, $p'_2 = p_2, \dots, p'_r = p_r$, $p'_{r+1} = p_1 + p_{r+1}$. То есть проекция максимального сечения множества B^λ на $r - 1$ -мерное пространство переменных есть то же множество B^λ на единицу меньшей размерности, относительно другого набора независимых переменных и вероятностей.

«Фрактальное» свойство множеств B^λ используется в технической лемме 17, где необходимо оценить разность объемов проекции множества B^λ и проекции множества B^1 . Благодаря (17) мы можем использовать готовый результат, полученный Т. Ридом.

Технические леммы 16-19 предваряют подраздел 2.2.3, в котором описывается собственно процесс редукции. В основе этого этапа лежит **Теорема 11** *Выражение*

$$[S_1(\sqrt{n}x_l + np_l)\phi(\mathbf{x})]_{\lambda_l(x^*)}^{\theta_l(x^*)}$$

можно переписать в виде

$$d[S_1(\sqrt{n}\theta_l(x^*) + np_l) - S_1(\sqrt{n}\lambda_l(x^*) + np_l)] + O\left(n^{-\frac{1}{2}}\right),$$

где

$$d = \frac{1}{e^{\frac{c}{2}} \sqrt{(2\pi)^r \prod_{j=1}^k p_j}}.$$

Доказательство этой теоремы отличается от доказательства аналогичного утверждения главы I, и, как уже отмечалось выше, нам необходимо воспользоваться свойствами гладкой функции ϕ . По завершении этапа мы можем сформулировать основное утверждение раздела 2.2.

Теорема 7 *Член J_2 можно представить в следующем виде*

$$J_2 = dn^{-\frac{r}{2}}(N^\lambda - n^{\frac{r}{2}}V^\lambda) + O(n^{-1}).$$

В разделе 2.3 исследуется применимость теоремы 2 к множествам B^λ . Леммы 21 и 22 подраздела 2.3.1 аналогичны рассмотренным в части I. Ключевой для последующих рассуждений является лемма 23 подраздела 2.3.2, в которой формулируются условия, которые мы должны потребовать от B^λ , чтобы не только применить теорему, но и должным образом оценить константу $I(n)$. Доказательство этой леммы представляло особую сложность, поскольку необходимо было вернуться к оригинальной работе Э. Главки [9] и дополнить изложенные в ней рассуждения.

Лемма 23 *Пусть для множеств $B(n)$ выполнены условия теоремы 2, и, кроме того,*

1. в любой точке границы множества гауссова кривизна $K_n(u)$ заключена в пределах, не зависящих от n, u , и равномерно по этим параметрам отделена от 0

$$0 < K_0 \leq K_n(u) \leq K_1,$$

2. для любого u на единичной сфере опорная функция $H_n(u)$ множества $B(n)$ ограничена равномерно по n и равномерно отделена от 0, то есть

$$H_1 \geq H_n(u) \geq H_0 > 0, \quad |u| = 1.$$

3. Частные производные $H_n(u)$ любого порядка равномерно по n ограничены сверху (на единичной сфере $|u| = 1$).

Тогда

$$|N - n^{\frac{r}{2}} V| = c \cdot n^{\frac{r}{2}-1+\frac{1}{r+1}},$$

где константа c не зависит от n .

Подраздел 2.3.3 содержит ряд технических утверждений, проверяющих условия леммы 23 в контексте множеств $B^\lambda(n)$. Используется терминология и утверждения из дифференциальной геометрии и оптимального управления. Следующие две леммы доказываются во многом аналогично главе I.

Лемма 27 Множества $B^\lambda(n)$ имеют в пространстве $\mathbb{R}^r$ однозначную гладкую параметризацию.

Лемма 28 Имеет место равномерная (по полярному углу θ) сходимость $r_n(\theta) \rightrightarrows r(\theta)$, а также аналогичная равномерная сходимость частных производных любого порядка.

Следствием последней леммы является то, что гауссова кривизна множеств $B^\lambda(n)$ сходится по n к гауссовой кривизне множества B^1 . Поскольку гауссова кривизна множества B^1 равномерно отделена от 0, начиная с некоторого N гауссова кривизна множеств $B^\lambda(n)$ также отделена от 0 равномерно по своему параметру.

В дальнейших рассуждениях используется известная из теории оптимального управления мера Хаусдорфа расстояния между множествами $haus(A, B)$. Доказав равномерную сходимость к 0 величины $haus(B^\lambda(n), B^1)$, мы приходим к лемме 29.

Лемма 29 Опорные функции $H_n(u)$ многообразий $B^\lambda(n)$ равномерно ограничены и равномерно отделены от 0 на единичной сфере $|u| = 1$.

Для того чтобы удовлетворились все требования леммы 23 еще необходимо показать, что все частные производные функции $H_n(u)$ равномерно по n ограничены сверху на единичной сфере. Это сделано в лемме 30. Окончательно получаем, что справедлива

Теорема 12 Для последовательности множеств $B^\lambda(n)$ выполнены все требования леммы 23.

В заключение раздел 2.4 еще раз перечисляет, в какой последовательности нужно использовать все полученные результаты, чтобы получить оценку (16).

Автор выражает благодарность своему научному руководителю Ульянову В. В. за постановку задачи и постоянное внимание к работе.

Итог

Получены два основных утверждения. Для любой размерности $r \geq 3$ имеем оценку скорости сходимости

$$\Pr(T_\lambda(\mathbf{X}) < c) = K_r(c) + O\left(n^{-1+\frac{1}{r+1}}\right).$$

Для случая трех ячеек группировки ($r = 2$) имеем

$$\Pr(T_\lambda(\mathbf{X}) < c) = K_2(c) + O\left(n^{-100/146}(\log n)^{315/146}\right).$$

Обе оценки получены с помощью сведения исходной задачи оценки скорости сходимости к известной из теории чисел задаче приближения числа точек решетки в выпуклом множестве его нормированным объемом.

Заключение

В диссертационной работе исследуется скорость сходимости семейства степенных статистик согласия к хи-квадрат распределению. Рассматривается как случай на плоскости, так и общий случай без наложения ограничений на размерность. Подтверждена непосредственная связь задачи оценивания скорости слабой сходимости степенных статистик согласия с известной задачей из теории чисел о приближении числа точек решетки в заданном выпуклом множестве его нормированным объемом; впервые указан алгоритм, посредством которого эта связь может быть использована для получения оценок скорости сходимости. Полученные в работе оценки справедливы для произвольных степенных статистик согласия; при этом они сопоставимы, а в некоторых случаях превосходят по точности имеющиеся результаты для статистики хи-квадрат.

Результаты работы, безусловно, очерчивают перспективу возможных дальнейших исследований на пути к получению полнофункциональных статистических критериев, использующих различные статистики семейства. Одно из возможных направлений — применение более точных результатов из теории чисел для уточнения оценок скорости сходимости к предельному распределению с частичным использованием наработанной техники. Также практический интерес представляет замена простой гипотезы на сложную или альтернативную

и последующий поиск скорости сходимости в этих случаях. Наконец, важным аспектом является оценка констант, входящих в ошибку аппроксимации предельным распределением.

Список публикаций автора по теме диссертации

- [1] АСЫЛБЕКОВ Ж. А., ЗУБОВ В. Н., УЛЬЯНОВ В. В. Асимптотические свойства почти квадратичных форм, Теория вероятн. и ее применен, т. 55, вып. 3, с. 617–618, 2010.
- [2] ЗУБОВ В. Н. Исследование скорости слабой сходимости степенных статистик согласия к хи-квадрат распределению, Обзорение прикладной и промышленной математики, т. 17, вып. 3, с. 365–366, 2010.
- [3] ASSYLBEKOV ZH. A., ULYANOV V. V., ZUBOV V. N. On the approximation of goodness-of-fit statistics for discrete three-dimensional data, Technical Report of the Hiroshima Statistical Research Group, 2008, no. 08-09, 26 p. <http://www.math.sci.hiroshima-u.ac.jp/stat/TR/TR08/TR08-09.PDF>
- [4] ULYANOV V. V., ZUBOV V. N. Refinement on the convergence of one family of goodness-of-fit statistics to chi-squared distribution. Hiroshima Mathematical Journal, **39**, 1, 133–161, 2009.
- [5] ULYANOV V. V., ZUBOV V. N. On the order of chi-squared approximation for the power divergence family of statistics, Preprint of SFB701, University of Bielefeld, 2010, no. 10-008, 48 p, www.mathematik.uni-bielefeld.de/sfb701/preprints/sfb08121.pdf
- [6] ZUBOV V. N. On the order of convergence of power-divergence goodness-of-fit statistics to chi-squared distribution, Abstracts of the 28th European Meeting of Statisticians, p. 364, University of Piraeus, Greece, August 17–22, 2010.