Исследование влияния вариантов группирования на мощность критериев согласия типа χ^2

Анна Д. Бушакова, Борис Ю. Лемешко
Новосибирский Государственный Технический Университет, г. Новосибирск, Российская Федерация
bushakova@ciu.nstu.ru

Аннотация **– Исследуется мощность критериев χ^2 Пирсона, Джапаридзе-Никулина при различных способах и числе интервалов группирования данных.**

Ключевые слова **– группирование, критерий согласия, мощность критерия**

I. Введение

В НАСТОЯЩЕЕ ВРЕМЯ в прикладных задачах статистического анализа в качестве моделей законов распределения вероятностей реальных случайных величин используется несколько десятков законов распределения. Остро стоит вопрос о правомерности применения того или иного закона распределения в качестве модели для описания наблюдаемых случайных величин.

При проверке согласия эмпирического закона распределения с теоретическим различают простые и сложные гипотезы. При простой проверяемой гипотезе вектор параметров закона распределения известен, а при проверке сложной параметры оцениваются. Чаще всего, параметры оцениваются по той же самой выборке, по которой проверяется гипотеза.

Случайность самой выборки предполагает возможное наличие ошибок в результатах статистических выводов. Различают ошибки двух видов. Ошибка первого рода заключается в отклонении верной проверяемой гипотезы H_0. Как правило, вероятность ошибки первого рода α задается. Ошибка второго рода заключается в принятии H_0, когда на самом деле справедлива некоторая конкурирующая гипотеза H_1. О вероятности ошибки второго рода β можно говорить только при задании конкурирующей гипотезы.

При наличии пары конкурирующих гипотез H_0 и H_1, задание α определяет величину β.

Величину $1-\beta$ называют мощностью критерия. Чем выше мощность критерия, тем лучше он различает конкурирующие гипотезы. Таким образом, если есть возможность выбора критерия, всегда следует предпочесть наиболее мощный.

В данной работе речь идет о мощности критериев типа χ^2. Применение такого рода критериев предусматривает разбиение области определения случайной величины на интервалы и подсчет количества наблюдений, попавших в них. Мощность критериев данного типа существенно зависит от способа разбиения на интервалы [1] и выбора их количества [2, 3].

В [4] показано, что мощность критерия χ^2 Пирсона определяется количеством сохранившейся при группировании фишеровской информации о параметрах закона, соответствующего проверяемой гипотезе H_0: чем меньше потери в информации Фишера, связанные с группированием выборки, тем выше мощность критерия по отношению к близким конкурирующим гипотезам.

Существует ряд критериев типа χ^2 для проверки гипотез о согласии эмпирического распределения с теоретическим. Рассмотрим критерий χ^2 Пирсона и критерий Джапаридзе-Никулина [5, 6].

II. Постановка задачи

Статистика критерия χ^2 Пирсона задается выражением:

$$\mathrm{X}^2(\theta) = n\sum_{i=1}^{k}\frac{n_i/n - p_i(\theta)}{p_i(\theta)}, \qquad (1)$$

где k – число интервалов группирования, n – количество наблюдений, n_i – количество наблюдений, попавших в i-й интервал, $p_i(\theta)$ – вероятность попадания в этот интервал. При справедливости H_0 в случае проверки простой гипотезы статистика в пределе подчиняется χ^2 распределению с числом степеней свободы $k-1$. А при проверке сложных и оценивании методом максимального правдоподобия s параметров – χ^2 распределению с числом степеней свободы $k-s-1$.

Критерий Джапаридзе-Никулина отличается от критерия Пирсона только при проверке сложных гипотез. Его статистика имеет вид [5]:

$$U^2(\theta)=\mathrm{X}^2(\theta)-nL^T(\theta)J^{-1}(\theta)L(\theta)\,, \quad (2)$$

$$L(\theta)=(l_1(\theta),...,l_s(\theta))\,,$$

$$l_j(\theta)=\sum_{i=1}^{k}\frac{n_i}{np_i(\theta)}\frac{\partial p_i(\theta)}{\partial\theta_j}\,,$$

где $J\ \theta$ – информационная матрица Фишера. При справедливости H_0 статистика подчиняется χ^2 распределению с числом степеней свободы $k-s-1$.

III. Теория

При справедливости конкурирующей гипотезы статистика критерия χ^2 Пирсона подчиняется нецентральному χ^2 распределению с тем же числом степеней свободы и параметром нецентральности:

$$\lambda=n\sum_{i=1}^{k}\frac{(p_i(\theta_1)-p_i(\theta))^2}{p_i(\theta)}\,. \quad (3)$$

При очень близких конкурирующих гипотезах (при малых $\delta\theta=\theta_1-\theta$ [4]) имеет место соотношение:

$$\lambda\approx n\delta\theta^T J_\Gamma(\theta)\delta\theta\,, \quad (4)$$

где $J_\Gamma(\theta)$ - информационная матрица Фишера по группированным данным.

Таким образом, чем меньше потери в информации Фишера, тем выше мощность критерия, выше его способность распознавать близкие конкурирующие гипотезы.

Естественно, что при любом группировании происходит потеря информации об исходном законе распределения. Это в свою очередь влияет на мощность критериев типа. Поэтому необходимо минимизировать потери, связанные с группированием.

В случае скалярного параметра задача асимптотически оптимального группирования, связанная с минимизацией потерь в информации Фишера решается однозначно [7].

В случае векторного параметра мы имеем дело с информационной матрицей. Тогда критерии минимизации потерь могут быть различными. В [7] задача асимптотически оптимального группирования была решена для случая D-оптимального группирования, когда в качестве минимизируемого функционала рассматривался определитель информационной матрицы по группированным данным

$$x^*=Arg\max_x\det(J_\Gamma(\theta))\,, \quad (5)$$

где под x понимается множество граничных точек k интервалов.

Однако далеко не очевидно, что в случае *D*-оптимального группирования мы имеем наилучший вариант, обеспечивающий максимальную мощность при близких конкурирующих гипотезах [8].

Отсюда возникает необходимость, во-первых, решения задач асимптотически оптимального группирования, обеспечивающих оптимум другим функционалам от информационной матрицы Фишера. Во-вторых, исследования мощности критериев при полученных вариантах асимптотически оптимального группирования.

Поэтому дополнительно рассмотрены две задачи асимптотически оптимального группирования. В задаче *A*-оптимального группирования максимизировался след

$$x^*=Arg\max_x sp(J_\Gamma(\theta))\,, \quad (6)$$

а в задаче *E*-оптимального группирования максимизировалось наименьшее собственное число матрицы $J_\Gamma(\theta)$ Фишера по группированным данным:

$$x^*=Arg\max_x\min_{i=1,s}\lambda_i(J_\Gamma(\theta))\,. \quad (7)$$

IV. Результаты экспериментов

Трудность решения задач (5)-(7) заключается в том, что они представляют собой многоэкстремальные задачи, а нас, естественно, интересует глобальный экстремум.

Задачи (6)-(7) *A*- и *E*-оптимального групппирования были решены для ряда законов распределения (нормального, логистического, Вейбулла, Коши, экстремальных значений) в виде, инвариантном относительно параметров. По аналогии с [7] были построены соответствующие таблицы асимптотически оптимального группирования.

Независимо от способа группирования элементы информационных матриц по группированным данным, соответствующие решениям задач (5)-(7), с ростом числа интервалов k сходятся к элементам информационной матрицы по негруппированным данным. В то же время, при конечных k различия в получаемых оптимальных решениях существенны.

Естественно, что существенным оказалось различие в условных распределениях статистик $G(X^2|H_1)$ при справедливости конкурирующей гипотезы H_1, и, соответственно, существенной оказалась разница в мощности критерия χ^2 Пирсона при использовании того или иного способа группирования. В качестве примера представлены асимптотически оптимальные граничные точки, соответствующие различным критериям оптимальности, при числе интервалов $k = 9$ для нормального закона, см. Табл. I. Вследствие симметричности приведены только 4 левые точки. В последней строке представлены граничные точки, которые обеспечивают максимум мощности критерия χ^2 Пирсона при проверке согласия с нормальным законом против конкурирующей гипотезы, соответствующей логистическому закону.

ТАБЛИЦА I
ОПТИМАЛЬНЫЕ ГРАНИЧНЫЕ ТОЧКИ

	k	x_1	x_2	x_3	x_4
A-optim	9	-2.3758	-1.6915	-1.1047	-0.4667
D-optim	9	-2.3188	-1.6218	-1.0223	-0.3828
E-optim	9	-1.8638	-1.1965	-0.6805	-0.2216
Optimum	9	-3.1616	-2.0856	-1.2676	-0.4601

На Рис. 1 показаны распределения статистики критерия χ^2 Пирсона в случае справедливости простой проверяемой гипотезы, соответствующей нормальному закону, и в случае справедливости конкурирующей гипотезы, соответствующей близкому логистическому закону, для числа интервалов $k = 9$ и объема выборки $n = 500$. Для сравнения здесь же показано распределение $G(X^2|H_1)$ в случае использования группирования, являющегося оптимальным для этих конкурирующих гипотез.

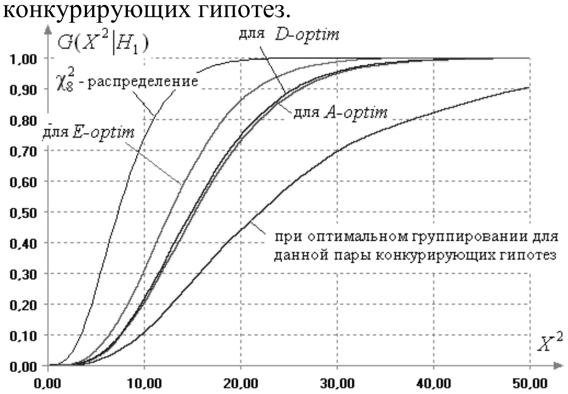

Рис. 1. Распределения статистик критерия χ^2 Пирсона при справедливости простой проверяемой гипотезы H_0, соответствующей нормальному закону, справедливости конкурирующей H_1, соответствующей логистическому, для $k = 9$ и $n = 500$

Мощность критерия χ^2 Пирсона исследовалась для различных пар конкурирующих гипотез и критериев оптимальности группирования. На Рис. 2 представлены зависимости мощности критерия χ^2 Пирсона от числа интервалов при различных вариантах оптимальности группирования при $n = 500$ для пары конкурирующих гипотез: H_0 – логистический закон; H_1 – нормальный закон (в случае проверки простой гипотезы).

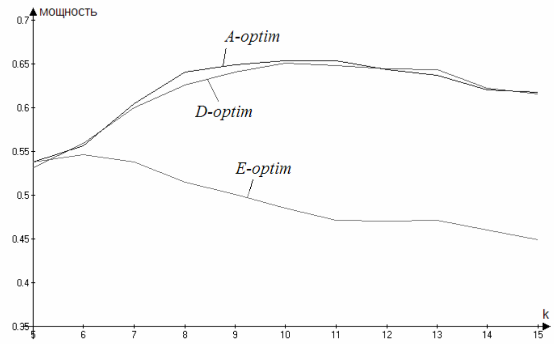

Рис. 2. Зависимость мощности критерия χ^2 Пирсона от числа интервалов для пары конкурирующих гипотез «логистический закон - нормальный»

Исследование распределений статистики критерия Джапаридзе-Никулина при справедливости проверяемой гипотезы H_0 показало, что оно очень быстро сходится к предельному χ^2-распределению с $k - s - 1$ степенями свободы [5]. Например, при проверке гипотез относительно законов логистического и нормального смоделированные распределения статистики Джапаридзе-Никулина хорошо согласуются с предельным уже при $n = 200$.

Мощность критерия Джапаридзе-Никулина исследовалась при A-, *D*-, E-оптимальном и равновероятном группировании.

На Рис. 3 показана зависимость мощности критериев χ^2 Пирсона и Джапаридзе-Никулина в зависимости от числа интервалов в случае применения *D*-оптимального группирования для пары конкурирующих гипотез: H_0 – нормальный закон; H_1 – логистический закон.

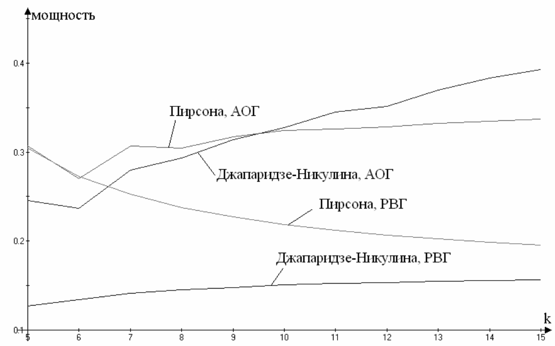

Рис. 3. Зависимость мощности критериев Пирсона и Джапаридзе-Никулина от числа интервалов при *D*-оптимальном группировании для пары конкурирующих гипотез «нормальный закон - логистический»

Для сравнения рассматривался еще один способ группирования, предложенный в [9] и который может использоваться при проверке гипотез при заданной конкурирующей гипотезе H_1. В случае, так называемых, интервалов Неймана-Пирсона [9] в качестве границ интервалов выбираются точки пересечения соответствующих плотностей. Исследования показали, что использование интервалов Неймана-Пирсона обеспечивает максимум мощности критерия относительно заданной H_1 при данном числе интервалов, по крайней мере, для критерия χ^2 Пирсона. Однако это не означает, что этот прием обеспечивает максимальную мощность для любого числа интервалов.

V. Обсуждение результатов

В случае исследования критерия χ^2 Пирсона, как правило, условные распределения статистик $G(X^2|H_1)$ при *A*-, *D*-оптимальном группировании близки. Соответственно, близки и мощности критериев при данных способах группирования. В случае же *E*-оптимального группирования мощность оказывается существенно ниже.

Критерий Джапаридзе-Никулина при малом числе интервалов уступает по мощности критерию χ^2 Пирсона (при проверке сложной гипотезы), а с ростом числа интервалов имеет заметное преимущество. При равновероятном способе группирования критерий Джапаридзе-Никулина существенно уступает по мощности критерию χ^2 Пирсона.

При *E*-оптимальном группировании мощность критерия Джапаридзе-Никулина ниже, чем в случае *A*- и *D*-оптимального группирования. Однако в отличие от схожести мощности критерия χ^2 Пирсона при этих вариантах группирования, мощность критерия Джапаридзе-Никулина при *A*-оптимальном группировании имеет свои особенности.

VI. Выводы и заключение

В данной работе были решены задачи *A*- и *E*-оптимального группирования для ряда законов распределения. Построены таблицы асимптотически оптимального группирования, которые могут быть использованы в задачах оценивания параметров по группированным данным и в критериях согласия.

Исследовано распределение статистик критериев при справедливости соответствующих проверяемых гипотез и их схождение к предельному.

Исследовано влияние способов группирования на мощность критерия χ^2 Пирсона и Джапаридзе-Никулина. Показано, что при *A*-оптимальном группировании мощность критерия χ^2 Пирсона несколько превосходит мощность его же в случае *D*-оптимального группирования. Применение *E*-оптимального группирования в критериях согласия неперспективно.

Исследована мощность критерия Джапаридзе-Никулина в зависимости от числа интервалов группирования. Отмечено, что при равновероятном группировании она существенно ниже, чем при асимптотически *A*- и *D*-оптимальном группировании. Критерий Джапаридзе-Никулина, целесообразно применять при большем количестве интервалов группирования

Список литературы

[1] Лемешко Б.Ю. Асимптотически оптимальное группирование наблюдений в критериях согласия // Заводская лаборатория, 1998. Т. 64. - №1. - С.56-64.

[2] Лемешко Б.Ю., Чимитова Е.В. О выборе числа интервалов в критериях согласия типа χ^2 // Заводская лаборатория. Диагностика материалов. 2003. Т. 69. – № 1. – С. 61-67.

[3] Лемешко Б.Ю., Чимитова Е.В. Максимизация мощности критериев типа χ^2 // Доклады Сибирского отделения Академии наук высшей школы. – Новосибирск, 2000. №2. С 53-61.

[4] Денисов В.И., Лемешко Б.Ю. Оптимальное группирование при обработке экспериментальных данных // Измерительные информационные системы. - Новосибирск, 1979. - С. 5-14.

[5] Джапаридзе К.О., Никулин М.С. Об одном видоизменении стандартной статистики Пирсона // Теория вероятностей и ее применения, 1974. т.19. № 4. С 886-888.

[6] Dzhaparidze K.O., Nikulin M.S. On the computation of chi-square-type statistics // Journal of Mathematical Sciences, 1995. – V.75. – № 5. – P.1910-1921.

[7] Денисов В.И., Лемешко Б.Ю., Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов: Монография. В 2-х ч. – Новосибирск: Новосиб. гос. техн. ун-т, 1993. – 347 с.

[8] Воинов В.Г. Об оптимальных свойствах критерия Рао-Робсон-Никулина // Заводская лаборатория. Диагностика материалов. 2006. – Т.72. – № 3. – С.65-70.

[9] Greenwood P.E., Nikulin M.S. A Guide to Chi-Squared Testing. New York: John Wiley & Sons, Inc. 1996. – 280 p.

Анна Дмитриевна Бушакова — Магистр, инженер-программист центра информатизации университета НГТУ.

Борис Юрьевич Лемешко — Профессор кафедры прикладной математики НГТУ, д.т.н., профессор.