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

Анна Д. Бушакова[1], Борис Ю. Лемешко[1]
[1] *Новосибирский Государственный Технический Университет*

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

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. Естественно, что это не покрывает все многообразия случайных величин. Вопрос о правомерности применения того или иного закона распределения в качестве модели для описания наблюдаемых случайных величин принимается на основании выводов с использованием критериев согласия.

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 и проверяется гипотеза.

Случайность самой выборки предполагает возможность ошибок в результатах статистических выводов. 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, что не всегда делается при использовании критериев согласия.

В случае полностью определенного критерия и наличия пары конкурирующих гипотез H_0 и H_1 задание α определяет и 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 критериев типа χ^2 предусматривает разбиение области определения случайной величины на интервалы и подсчет количества наблюдений, попавших в данные интервалы. Мощность критериев данного типа существенно зависит от способа разбиения на интервалы [1] и выбора количества интервалов [2, 3].

В [4] было 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.

II. РАССМАТРИВАЕМЫЕ КРИТЕРИИ

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

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 s параметров по группированным данным – 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 s - число параметров закона, оцениваемых методом максимального правдоподобия по негруппированным данным, $J\ \theta$ - информационная матрица Фишера. При справедливости H_0 статистика подчиняется χ^2-распределению с числом степеней свободы $k-s-1$.

III. АСИМПТОТИЧЕСИ ОПТИМАЛЬНОЕ ГРУППИРОВАНИЕ

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, необходимость исследования мощности критериев при полученных вариантах асимптотически оптимального группирования.

В данном случае дополнительно рассмотрены две постановки задач 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

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 группирования были решены для ряда законов распределения (нормального, логистического, Вейбулла, Коши, экстремальных значений) в виде, инвариантном относительно параметров закона. 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 различие 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 таблице 1 в качестве примера представлены асимптотически оптимальные граничные точки, соответствующие различным критериям оптимальности, при числе интервалов $k=9$ для нормального закона. В таблице вследствие симметричности приведены только 4 левые точки. В последней строке представлены граничные, которые обеспечивают максимум мощности критерия χ^2 Пирсона при проверке согласия с нормальным законом против конкурирующей гипотезы, соответствующей логистическому закону с плотностью

$$f(x)=\frac{\pi}{\theta_0\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\right]^2 .$$

ТАБЛИЦА 1
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

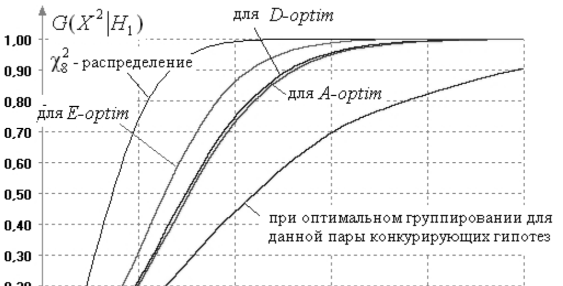

Рис. 1. Распределения статистик критерия χ^2 Пирсона при справедливости простой проверяемой гипотезы H_0, соответствующей нормальному закону, справедливости конкурирующей H_1, соответствующей логистическому, для $k=9$ и $n=500$

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 2-х конкурирующих гипотез.

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

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).

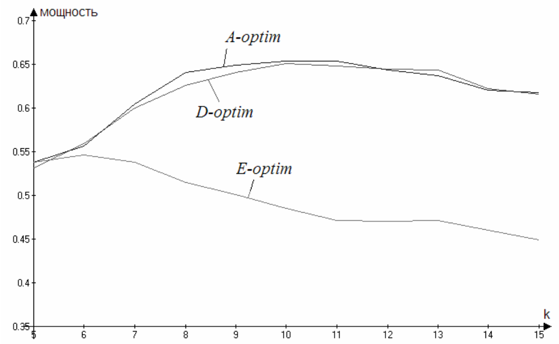

Рис. 2. Зависимость мощности критерия χ^2 Пирсона от числа интервалов для пары конкурирующих гипотез «логистический закон - нормальный»

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 D-оптимальном и равновероятном группировании.

На рис. 3 показана зависимость мощности критериев χ^2 Пирсона и Джапаридзе-Никулина от числа интервалов в случае применения D-

оптимального группирования (при $n = 200$) для пары конкурирующих гипотез: H_0 – нормальный закон; H_1 – логистический закон. Как видим, при малом числе интервалов критерий Джапаридзе-Никулина уступает по мощности критерию χ^2 Пирсона (при проверке сложной гипотезы), а с ростом числа интервалов имеет заметное преимущество. Поэтому критерий Джапаридзе-Никулина, целесообразно применять при большем количестве интервалов группирования. При равновероятном способе группирования критерий Джапаридзе-Никулина существенно уступает по мощности критерию χ^2 Пирсона.

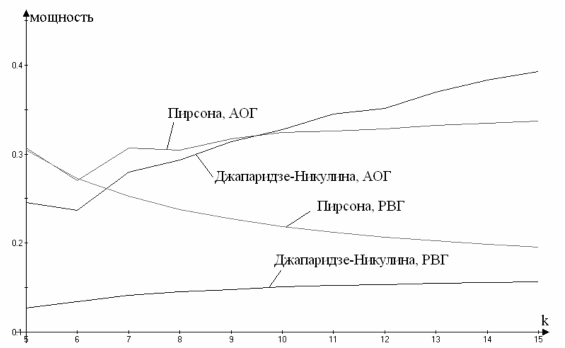

Рис. 3. Зависимость мощности критериев Пирсона и Джапаридзе-Никулина от числа интервалов при *D*-оптимальном группировании для пары конкурирующих гипотез «нормальный закон - логистический»

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. ЗАКЛЮЧЕНИЕ

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

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. Отмечено, что при равновероятном группировании мощность критерия Джапаридзе-Никулина существенно ниже, чем при асимптотически *A*- и *D*-оптимальном группировании.

Работа выполнена при частичной финансовой поддержке РФФИ (проект № 06-01-0059а).

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	Бакалавр прикладной математики и информатики (2007 г.), магистрант факультета прикладной математики и информатики НГТУ.
Борис Юрьевич Лемешко	Профессор кафедры прикладной математики НГТУ, д.т.н., профессор.