Материалы международной НТК "Информатика и проблемы телекоммуникаций". - Новосибирск, 1997. - С. 120-123.

О ВЛИЯНИИ СПОСОБА ГРУППИРОВАНИЯ ДАННЫХ НА РАСПРЕДЕЛЕНИЯ СТАТИСТИК χ^2 ПИРСОНА И ОТНОШЕНИЯ ПРАВДОПОДОБИЯ

Б.Ю.Лемешко, С.Н.Постовалов
Новосибирский государственный технический университет

Независимо от того, каким образом сгруппированы данные в интервалы, при использовании критериев χ^2 Пирсона и отношения правдоподобия в качестве предельных распределений соответствующих статистик используются χ^2-распределения. Факт оценивания по выборке параметров учитывается соответствующим уменьшением числа степеней свободы предельного χ^2-распределения, а вид закона $f(x,\theta)$ – использованием асимптотически оптимального группирования для него, гарантирующего максимальную мощность этих критериев при близких альтернативах [1].

Давно очевидно, что вычисленные по конкретной выборке значения рассматриваемых статистик очень сильно зависят от того, как сгруппированы данные: выбрали интервалы группирования одним способом – нулевая гипотеза H_0 о согласии должна быть отвергнута, другим – нет оснований её отвергать. Ясно, что ***предельные распределения*** статистик критериев χ^2 Пирсона и отношения правдоподобия ***зависят*** не только от числа оцененных параметров, но и ***от способа группирования***, вида исследуемого закона распределения $f(x,\theta)$ и оцениваемого параметра.

При практическом использовании критериев согласия выбирают либо интервалы равной длины, либо интервалы равной вероятности (равной частоты), либо асимптотически оптимальные интервалы [1]. Использование интервалов равной вероятности было предложено ещё Манном и Вальдом в 1942 г. [2]. Разбиение области определения случайной величины (размаха выборки) на интервалы равной длины неоднозначно. Более определенными способами являются равновероятное и асимптотически оптимальное группирование. При асимптотически оптимальном группировании лучше улавливаются небольшие отклонения выборки от предположений.

Целью данной работы явилось стремление выяснить как значительно отличаются законы распределения одной и той же статистики при равновероятном и асимптотически оптимальном группировании, насколько сильно законы распределения статистик отличаются от соответствующего предельного χ^2-распределения, в том числе в зависимости от того, сколько параметров оценивалось по выборке и с каким законом проверяется согласие?

По каждому закону распределения $f(x,\theta)$ моделировалась серия из $N=500$ выборок объемом $n=130$. Псевдослучайные величины имитировалась по методу обратных функций. В качестве датчика равномерно распределенных псевдослучайных чисел использовался стандартный датчик, реализованный в ***C++***. Оценки параметров находились по методу максимального правдоподобия по негруппированным данным.

Статистика χ^2 Пирсона вычисляется в соответствии с соотношением

$$S_{\chi^2} = n\sum_{i=1}^{k}\frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)},$$

где n_i – количество наблюдений, попавших в интервал, $P_i(\theta)$ - вероятность попадания в i-й интервал, и при истинной гипотезе H_0 в пределе подчиняется (*должна*) χ^2_{k-r-1}-распределению, где $k-r-1$ - число степеней свободы, k - число интервалов, r - количество оцененных по выборке параметров.

Статистика критерия отношения правдоподобия [2]

$$S_{\text{îï}} = -2\ln l = -2\sum_{i=1}^{k} n_i \ln\left(\frac{P_i(\theta)}{n_i/N}\right)$$

имеет те же асимптотические распределения.

Результаты моделирования и последующего анализа указывают на то, что предельные распределения статистик S_{χ^2}, $S_{\text{îï}}$ существенно отличаются при различных способах группирования. При этом эмпирический закон распределения статистики при асимптотически оптимальном группировании обычно ближе к теоретическому χ^2-распределению, чем при равновероятном группировании. Мало того, распределения статистик зависят не только от количества оцененных по выборке параметров, но и от того, какой параметр оценивался. Например, оценивание параметра сдвига приводит к более значительному изменению распределения статистики, чем оценивание масштабного параметра.

χ^2_l-распределение с числом степеней свободы l является частным случаем гамма-распределения с основным параметром, равным $l/2$, и масштабным - 0.5. Оценивание одного из параметров наблюдаемого закона учитывается уменьшением числа степеней свободы на 1. На самом деле, если измерять изменение предельного закона “в степенях свободы”, то оказывается, что оценивание даже параметра сдвига обычно приводит к изменению “числа степеней” на величину меньшую 1, еще к меньшему изменению в степенях свободы приводит оценивание масштабного параметра. Это особенно заметно при оценивании 2-х параметров. Всё это говорит о том, что уменьшение числа степеней свободы на r оцененных параметров и использование в критерии согласия χ^2_{k-r-1}-распределения чревато занижением вероятности вида $P\{S > S^*\}$ и, следовательно, риском отвергнуть верную гипотезу H_0. Можно

сделать предварительный вывод, что особенно существенно занижение $P\{S > S^*\}$ при меньшем числе интервалов группирования.

На рис. 1 для иллюстрации представлены результаты моделирования статистик при 5 интервалах группирования и оценивании 2-х параметров нормального распределения. На рисунке приведена теоретическая функция χ^2_2-распределения (*а*), эмпирические функции статистики S_{χ^2} при асимптотически оптимальном (*б*) и равновероятном группировании (*в*).

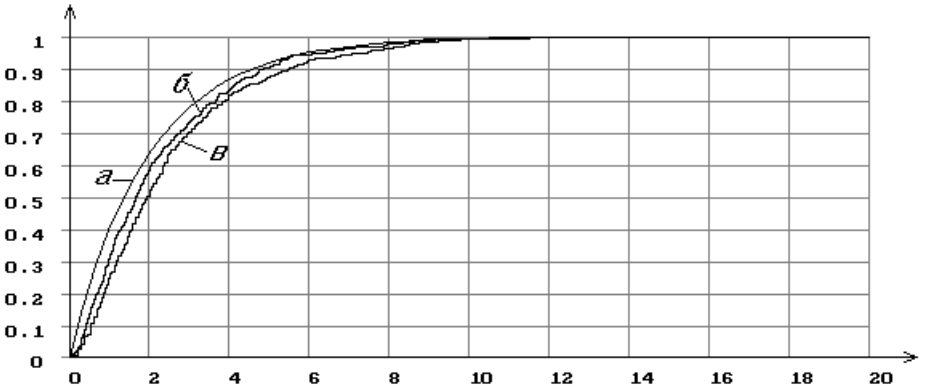

Рис. 1. Распределения статистики S_{χ^2} при 5 интервалах группирования и оценивании 2-х параметров нормального закона

При идентификации типов предельных законов распределения статистик оказалось, что эмпирические законы распределения статистик критериев с достаточно высокой степенью точности практически всегда описываются *гамма-распределением*.

В качестве предварительного вывода можно высказать предположение, что эмпирические распределения статистики отношения правдоподобия оказываются ближе к предельному теоретическому χ^2_{k-r-1}-распределению, чем соответствующие распределения статистики критерия χ^2 Пирсона.

Результаты моделирования и анализа, позволяют сделать следующие выводы:

1. Предельные распределения статистик критериев отношения правдоподобия и χ^2 Пирсона существенно зависят от способа группирования.
2. При этом эмпирический закон распределения статистики при асимптотически оптимальном группировании ближе к теоретическому χ^2_{k-r-1}-распределению, чем при равновероятном группировании.
3. Распределения статистик зависят не только от количества оцененных по выборке параметров, но и от того, какой параметр оценивался. Оценивание параметра сдвига приводит к более значительному изменению распределения статистики, чем оценивание масштабного параметра.
4. В целом, *при оценивании r параметров число степеней свободы предельного распределения уменьшается на "число степеней свободы" $< r$.*

1. Денисов В.И., Лемешко Б.Ю., Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов: В 2 ч. / Новосиб. гос. техн. ун-т. - Новосибирск, 1993. - 346 с.
2. Кендалл М., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. - 900 с.