СБОРНИК НАУЧНЫХ ТРУДОВ НГТУ. - 1997. - № 1(6). - 23-32
УДК 519.24

К ВОПРОСУ О РАСПРЕДЕЛЕНИЯХ СТАТИСТИК НЕПАРАМЕТРИЧЕСКИХ КРИТЕРИЕВ СОГЛАСИЯ

Б.Ю. ЛЕМЕШКО♣, С.Н. ПОСТОВАЛОВ♥

На основании результатов моделирования показано, что в случае оценивания по выборке параметров предельные распределения статистик непараметрических критериев согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса при справедливой гипотезе H_0 настолько сильно отличаются соответственно от законов $K(s)$, χ^2_2, $a1(s)$ и $a2(s)$, что последние ни в коем случае не должны использоваться в такой ситуации. Для ряда законов распределения случайных величин идентифицированы законы распределения статистик непараметрических критериев при различном количестве оцененных параметров. Полученные законы при практическом использовании критериев согласия позволят делать более надежные статистические выводы.

Наиболее часто в практике статистического анализа с необходимостью использования критериев согласия приходится сталкиваться после оценивания по этой же выборке параметров предполагаемого закона распределения. К сожалению, *в этом случае предельные распределения* статистик таких непараметрических критериев, как критерии Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, в случае справедливости нулевой гипотезы вида H_0: $f(x,\theta_0) = f(x,\hat{\theta})$, где $f(\cdot)$ - плотность распределения наблюдаемого закона; θ_0 - истинное значение параметра; $\hat{\theta}$ - оценка параметра, вычисленная по выборке, *отличаются* от ситуации, когда по выборке не оцениваются параметры. На самом деле предельные распределения зависят как от числа оцененных параметров, так и от вида исследуемого закона распределения $f(x,\theta)$.

На практике при использовании непараметрических критериев согласия факт зависимости предельного распределения от оценивания параметров по разным причинам обычно не учитывается: берется

♣ Доцент кафедры прикладной математики, канд. техн. наук
♥ Аспирант кафедры прикладной математики

предельное распределение статистики, как будто параметры и не оценивались. Это приводит к сильно завышенным значениям вероятностей "согласия" вида $P\{S > S^*\}$, где S^* - значение статистики, вычисленное по выборке. Как сильно мы ошибаемся, если, используя непараметрический критерий согласия, не учитываем факт оценивания по выборке параметров и вид предполагаемого закона распределения? Конечно, желательно точно знать предельные распределения этих статистик в зависимости от того, сколько параметров оценивалось по выборке и с каким законом проверяется согласие. Очевидно, что теоретически найти решение этой задачи для множества законов, используемых для описания реальных величин, очень сложно.

Один из выходов нам видится в моделировании эмпирических законов распределения статистик непараметрических критериев и в последующей идентификации этих законов. В данной статье мы приводим сводные результаты моделирования и анализа, которые, с нашей точки зрения, могут с успехом применяться при решении практических задач проверки гипотез о согласии с использованием непараметрических критериев после вычисления оценок параметров распределения по той же выборке.

Статистики Колмогорова и Смирнова определяется соответственно выражениями [1]

$$S_k = \frac{(6nD_n + 1)^2}{18n} \quad \text{и} \quad S_m = \frac{(6nD_n^+ + 1)^2}{9n},$$

где

$$D_n = \max(D_n^+, D_n^-),\ D_n^+ = \max_{1 \le i \le n}\left\{\frac{i}{n} - F(x_i)\right\},\ D_n^- = \max_{1 \le i \le n}\left\{F(x_i) - \frac{i-1}{n}\right\},$$

n - объем выборки; $x_1, x_2, \ldots, x_n$ - упорядоченные по возрастанию выборочные значения; $F(x)$ - функция распределения, согласие с которой проверяется. Распределение величины $\sqrt{S_k / 2}$, если по выборке не оценивались параметры, в пределе подчиняется закону Колмогорова с функцией распределения $K(x)$ [1]. Гипотеза о согласии не отвергается, если

$$P\{S_k > S_k^*\} = 1 - K\left(\sqrt{\frac{S_k^*}{2}}\right) > \alpha.$$

В аналогичной ситуации статистика Смирнова S_m подчиняется в пределе распределению χ^2 с числом степеней свободы, равным 2. Гипотеза о согласии не отвергается, если

$$P\{S_m > S_m^*\} = \int_{S_m^*}^{\infty} \frac{1}{2} e^{-x/2} dx = 1 - e^{-S_m^*/2} > \alpha .$$

Статистики Мизеса имеют вид [1]

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left\{ F(x_i) - \frac{2i-1}{2n} \right\}^2$$

и

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n} \left\{ \frac{2i-1}{2n} \ln F(x_i) + \left(1 - \frac{2i-1}{2n}\right) \ln(1 - F(x_i)) \right\} .$$

Для этих статистик также известны предельные распределения вероятностей [1]

$$\lim_{n\to\infty} P\{n\omega_n^2 < x\} = a1(x) , \quad \lim_{n\to\infty} P\{n\Omega_n^2 < x\} = a2(x) .$$

Гипотезы о согласии не отвергаются, если выполняются неравенства

$$P\{S_\omega > S_\omega^*\} = 1 - a1(S_\omega^*) > \alpha \quad \text{и} \quad P\{S_\Omega > S_\Omega^*\} = 1 - a2(S_\Omega^*) > \alpha .$$

В работе моделировались выборки статистик $\sqrt{S_k/2}$, S_m, S_ω, S_Ω.

На рис.1 приведены результаты моделирования величины $\sqrt{S_k/2}$, используемой в критерии Колмогорова, при проверке гипотез о согласии с нормальным распределением при справедливости гипотезы H_0. На рис. 1 - 4 представлены эмпирические функции распределения статистики, когда по выборке не оценивались параметры (*а*), по выборке оценивались только масштабный параметр (*б*) (в данном случае σ), параметр сдвига (*в*) (в данном случае μ) и одновременно оба параметра (*г*). Здесь же приведена функция распределения Колмогорова (*д*), которому подчиняется статистика $\sqrt{S_k/2}$, если по выборке не оцениваются параметры. Результаты проверки согласия эмпирического распределения (*а*) с распределением Колмогорова (*д*) очень хорошее. В то же время весьма наглядно отличие эмпирических функций распределения *б, в, г* от распределения Колмогорова (*д*). Это отличие позволяет судить о величине тех ошибок, которые мы допус-

каем, не учитывая факта оценивания параметров конкретного распределения при использовании критерия Колмогорова.

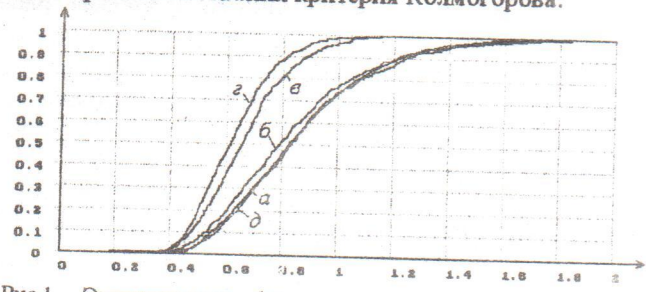

Рис.1. Эмпирические функции распределения статистики $\sqrt{S_k/2}$ Колмогорова при различном количестве оцениваемых параметров нормального закона

Аналогичная картина распределения статистики Колмогорова при справедливой гипотезе H_0 наблюдается для распределения Лапласа на рис. 2 и для распределения Коши на рис. 3. Результаты, полученные при моделировании непараметрических статистик, однозначно указывают на то, что предельные распределения статистик непараметрических критериев при условии оценивания параметров конкретного закона настолько сильно отличаются соответственно от распределений Колмогорова, χ^2_2, $a1(s)$ и $a2(s)$, что использование последних никак не может быть оправдано из-за высокого риска неверных выводов.

По каждому закону распределения $f(x,\theta)$ моделировалась серия из $N = 500$ выборок объемом $n = 130$. Оценки параметров находились по методу максимального правдоподобия по негруппированным данным.

При идентификации типов предельных законов распределения непараметрических статистик в зависимости от вида закона наблюдаемой случайной величины и количества оцениваемых по наблюденной выборке параметров использовалось множество законов и семейств распределений, включенных в программную систему [2].

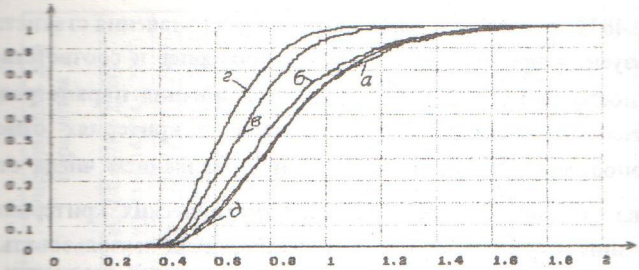

Рис.2. Эмпирические функции распределения статистики $\sqrt{S_k/2}$ Колмогорова при различном количестве оцениваемых параметров распределения Лапласа

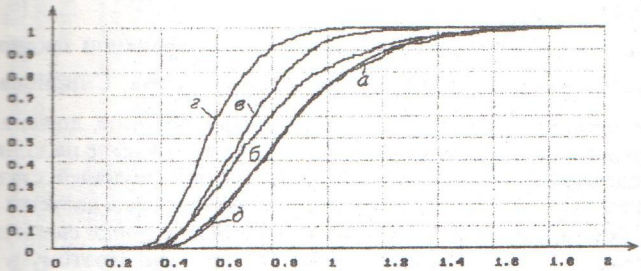

Рис.3. Эмпирические функции распределения статистики $\sqrt{S_k/2}$ Колмогорова при различном количестве оцениваемых параметров распределения Коши

Оказалось, что почти всегда с достаточно высокой степенью точности эмпирические законы распределения статистик непараметрических критериев описываются одним из двух законов распределения: *логарифмически нормальным* или *гамма-распределением*.

На рис. 4 представлены результаты выравнивания распределения статистики Колмогорова при оценивании одновременно двух параметров нормального распределения, отражены эмпирическая функция распределения статистики и функция распределения логарифмически нормального распределения с параметрами

$\mu = -0.4879$, $\sigma = 0.2235$. Здесь же приведены значения статистик всех используемых при проверке согласия критериев и соответствующие вероятности вида $P\{S > S^*\}$. Факт оценивания параметров логарифмически нормального распределения в критериях отношения правдоподобия и χ^2 Пирсона учтен уменьшением числа степеней свободы χ^2-распределения. В непараметрических критериях факт оценивания параметров не учитывался. Если использовать полученные результаты для предельных распределений статистик, то вероятность вида $P\{S > S^*\}$ для критерия Колмогорова составит 0.3111, а не 0.8830, для критерия Смирнова - 0.3505, а не 0.5039, для критерия ω^2 Мизеса - 0.4420, а не 0.8543, для критерия Ω^2 Мизеса - 0.3837, а не 0.8669.

В табл. 1 - 4 сведены результаты идентификации законов соответственно для статистик критериев Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса. Информация, представленная в таблице, должна интерпретироваться следующим образом. Указание в клетке на конкретное распределение означает, что выборка соответствующей статистики наиболее хорошо описывается данным законом (согласуется с законом). Без оценивания параметров выборки *хорошо согласуются* с предельными законами, определенными теорией. Поэтому в первом столбце таблиц указывается лишь закон, с которым лучше всего согласуется выборка. В случае если согласие с каким-то законом не очень хорошее (гипотеза о согласии принимается с уровнем значимости $\alpha = 0.1 \div 0.05$), то соответствующий закон указан на сером фоне. В таблицах через $\ln N(\mu, \sigma)$ обозначено логарифмически нормальное распределение с функцией плотности

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2},$$

через $\gamma(\theta_0, \theta_1, \theta_2)$ - гамма-распределение с функцией плотности

$$f(x) = \frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)} (x - \theta_2)^{\theta_0 - 1} e^{-\theta_1 (x - \theta_2)}.$$

Таблица 1

Предельные распределения статистики Колмогорова

Распределение случайной величины	Параметры по выборке не оценивались	Оценивался только масштабный параметр	Оценивался только параметр сдвига	Оценивалось два параметра
Нормальное	$K(x)$, lnN(-0.1808,0.3060)	lnN(-0.2245,0.3157) γ(3.1875,6.2605,0.3312)	lnN(-0.4248,0.2350) γ(3.5392,11.051,0.3520)	lnN(-0.4879,0.2235) γ(5.4165,16.024,0.2917)
Коши	lnN(-0.1808,0.3060)	γ(2.6218,5.7693,0.3318)	γ(4.4365,11.101,0.3100)	γ(3.7771,12.507,0.2909)
Лапласа	lnN(-0.1808,0.3060)	γ(3.4375,6.5969,0.3056)	γ(3.3433,9.3957,0.3532)	γ(4.3667,12.585,0.2945)
Экспоненциальное	lnN(-0.1808,0.3060)	lnN(-0.3324,0.2545)		
Полунормальное	lnN(-0.1808,0.3060)	lnN(-0.2956,0.2684)		
Рэлея	lnN(-0.1808,0.3060)	lnN(-0.3324,0.2545)		

Таблица 2

Предельные распределения статистики Смирнова

Распределение случайной величины	Параметры по выборке не оценивались	Оценивался только масштабный параметр	Оценивался только параметр сдвига	Оценивалось два параметра
Нормальное	γ(0.9564,0.4724,0)	γ(0.7737,0.4269,0.0024)	lnN(0.2761,0.5533)	lnN(0.1051,0.5478)
Коши	γ(0.9564,0.4724,0)	γ(0.7782,0.4814,0.0009)	γ(1.3746,0.9748,0.0213)	γ(1.3257,1.3842,0.0149)
Лапласа	γ(0.9564,0.4724,0)	γ(0.7744,0.4407,0.0021)	γ(1.4691,1.0715,0.0864)	lnN(-0.1539,0.8078)
Экспоненциальное	γ(0.9564,0.4724,0)	lnN(0.1983,0.7328)		
Полунормальное	γ(0.9564,0.4724,0)	γ(1.2931,0.8505,0.1104)		
Рэлея	γ(0.9564,0.4724,0)	lnN(0.1983,0.7328)		

Таблица 3

Предельные распределения статистики ω^2 Мизеса

Распределение случайной величины	Параметры по выборке не оценивались	Оценивался только масштабный параметр	Оценивался только параметр сдвига	Оценивалось два параметра
Нормальное	lnN(-2.0840,0.7948)	lnN(-2.2232,0.8366)	lnN(-2.7759,0.5739)	lnN(-3.0021,0.5195)
Коши	lnN(-2.0840,0.7948)	lnN(-2.3380,0.8893)	γ(1.4561,19.8471,0.016)	lnN(-3.0212,0.6632)
Лапласа	lnN(-2.0840,0.7948)	lnN(-2.2463,0.8487)	γ(1.5683,24.247,0.0157)	lnN(-2.9690,0.5576)
Экспоненциальное	lnN(-2.0840,0.7948)	lnN(-2.5730,0.6231)		
Полунормальное	lnN(-2.0840,0.7948)	lnN(-2.4650,0.6642)		
Рэлея	lnN(-2.0840,0.7948)	lnN(-2.5730,0.6231)		

Таблица 4

Предельные распределения статистики Ω^2 Мизеса

Распределение случайной величины	Параметры по выборке не оценивались	Оценивался только масштабный параметр	Оценивался только параметр сдвига	Оценивалось два параметра
Нормальное	lnN(-0.2229,0.6838)	lnN(-0.4036,0.7185)	lnN(-0.8159,0.5266)	lnN(-1.0973,0.4518)
Коши	lnN(-0.2229,0.6838)	lnN(-0.4464,0.7398)	γ(1.4191,2.8391,0.1457)	lnN(-0.9311,0.6138)
Лапласа	lnN(-0.2229,0.6838)	lnN(-0.4259,0.7253)	γ(1.6593,3.7008,0.1411)	lnN(-0.9670,0.5033)
Экспоненциальное	lnN(-0.2229,0.6838)	γ(1.7812,3.9758,0.1356)		
Полунормальное	lnN(-0.2229,0.6838)	γ(1.6100,3.2839,0.1472)		
Рэлея	lnN(-0.2229,0.6838)	γ(1.7812,3.9758,0.1356)		

В данной работе исследовались распределения статистик, которые представлены в таблицах, когда наблюдаемые случайные величины распределены в соответствии с законами: нормальным, Коши, Лапласа, экспоненциальным, полунормальным, Рэлея.

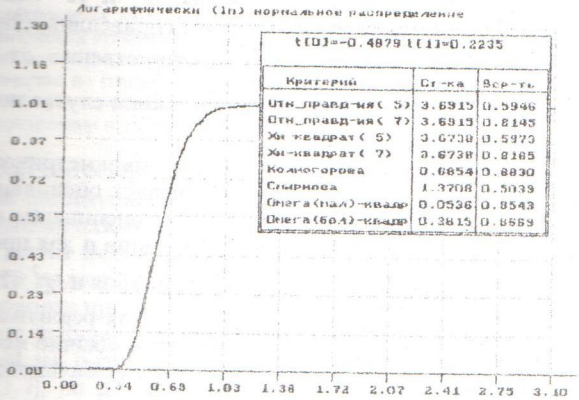

Рис.4. Эмпирическая функция статистики Колмогорова при справедливой гипотезе H_0 и оценивании 2-х параметров нормального распределения и выравнивающая её функция распределения логарифмически нормального распределения

В заключение посмотрим, что будет получаться, если мы будем использовать распределение Колмогорова для вычисления вероятности вида $P\{S > S^*\}$ в случае, когда по выборке предварительно вычисляются оценки параметров нормального распределения. Распределение статистики Колмогорова в этом случае хорошо описывается логарифмически нормальным $\ln N(-0.4879, 0.2235)$. Для распределения Колмогорова $P\{S > 0.9\} = 0.392731$, а для логарифмически нормального - $P\{S > 0.9\} = 0.043498$. Это означает, что при значении статистики $S^* = 0.9$ по распределению Колмогорова мы, не задумываясь, примем гипотезу H_0, когда на самом деле даже при уровне значимости $\alpha = 0.05$ она должна быть отклонена.

ЗАКЛЮЧЕНИЕ

На основании проведенных исследований можно сделать следующие выводы:

1. Предельные распределения статистик непараметрических критериев согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса при оценивании по выборке параметров в случае справедливости гипотезы H_0 настолько сильно отличаются соответственно от законов $K(s)$, χ^2_2, $a1(s)$ и $a2(s)$, что последние ни в коем случае не должны использоваться в такой ситуации.

2. На предельные распределения всех непараметрических статистик наиболее значительное влияние оказывает оценивание параметра сдвига, в существенно меньшей степени - оценивание масштабного параметра. Кстати, этот же вывод справедлив и для предельных распределений статистик отношения правдоподобия и χ^2 Пирсона.

3. Достаточно хорошая аппроксимация для реальных распределений статистик непараметрических критериев обычно может быть получена с использованием логарифмически нормального распределения и/или гамма-распределения.

4. Для ряда законов распределения случайных величин идентифицированы законы распределения статистик непараметричеких критериев при различном количестве оцененных параметров. Полученные законы при практическом использовании критериев согласия позволят делать более надежные статистические выводы.

[1] БОЛЬШЕВ Л.Н., СМИРНОВ Н.В. *Таблицы математической статистики*. - М.: Наука, 1983. - 416 с.

[2] ЛЕМЕШКО Б.Ю. *Статистический анализ одномерных наблюдений случайных величин: Программная система*. - Новосибирск: Изд-во НГТУ, 1995. - 125 с.