УДК 519.2

Моделирование распределений статистик непараметрических критериев согласия при потере свойства "свободы от распределения"

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

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, при 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. При оценивании по выборке параметров рассматриваемые критерии теряют свойство "свободы от распределения", и предельные распределения статистик на самом деле зависят как от числа оцененных параметров, так и от вида исследуемого закона распределения $f(x,\theta)$. Распространенная ошибка, связанная с пренебрежением этим фактом, чаще всего приводит к необоснованному принятию нулевой гипотезы, что подчеркивается в работах [1,2], из-за сильно завышенных значений вероятностей "согласия" вида $P\{S>S^*\}$, где S^* - значение статистики, вычисленное по выборке.

Если объем выборки достаточно велик, можно, опираясь на результаты, полученные в [3], оценивать параметры распределения по одной половине выборки, а проверять согласие по другой. В такой ситуации применение предельных распределений рассматриваемых критериев вполне обосновано. Но в этом случае как при оценивании, так и при проверке согласия мы используем только половину имеющейся информации, что, естественно, сказывается на качестве статистических выводов. К тому же, объемы выборок, зачастую имеющиеся в распоряжении исследователя, не настолько велики, чтобы можно было смириться с потерей части информации при оценивании параметров.

Предельные распределения статистики критерия ω^2 Мизеса при оценивании одного из двух или обоих параметров нормального распределения подробно исследованы в [4], где приведены их таблицы. Процентные точки для модифицированных статистик критериев Колмогорова (типа Колмогорова) в такой ситуации представлены, например, в [1]. Там же для обоих критериев приведены процентные точки при проверке экспоненциальности и оценивании его неизвестного масштабного параметра.

Очевидно, что теоретически найти решение задачи определения предельных распределений непараметрических статистик для множества законов, используемых для описания реальных величин, очень сложно. Наиболее практич-

ный выход нам видится в моделировании предельных законов распределения статистик непараметрических критериев и в последующей идентификации полученных эмпирических законов.

В данной работе моделировались распределения статистик Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса. Но мы приведем только результаты моделирования распределений статистики Смирнова. Статистика Смирнова определяется выражением [5]

$$S_m = \frac{(6nD_n^+ + 1)^2}{9n},$$

где $D_n^+ = \max\limits_{1 \le i \le n}\left\{\frac{i}{n} - F(x_i)\right\}$, 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 S_m подчиняется в пределе распределению χ^2 с числом степеней свободы, равным 2. Гипотеза о согласии не отвергается, если

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

На рис. 1 представлены результаты моделирования статистики Смирнова S_m при справедливой гипотезе H_0, соответствующей нормальному распределению, на рис. 2 - соответствующей логистическому распределению, на рис. 3 - распределению Лапласа, на рис. 4 - распределению Коши, а на рис. 5 - экспоненциальному распределению. На этих рисунках *0* помечена функция распределения χ_2^2.

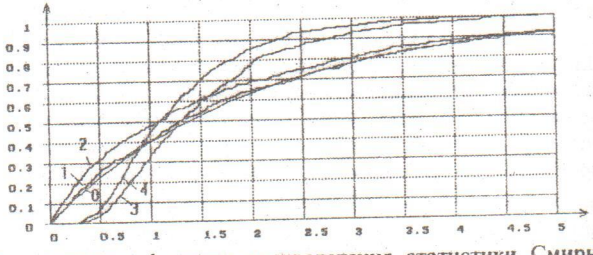

Рис.1. Эмпирические функции распределения статистики Смирнова при различном количестве оцениваемых параметров нормального закона: *0* - функция распределения χ_2^2 (с двумя степенями свободы); *1* - по выборке не оценивались параметры; *2* - по выборке оценивался только масштабный параметр; *3* - оценивался только параметр сдвига; *4* - оценивались одновременно оба параметра;

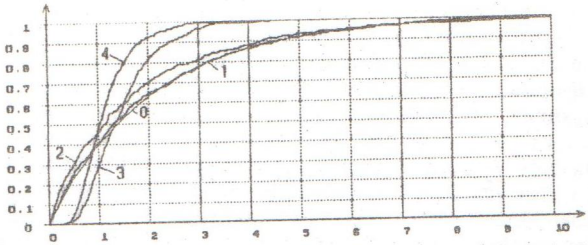

Рис.2. Эмпирические функции распределения статистики Смирнова при различном количестве оцениваемых параметров логистического распределения

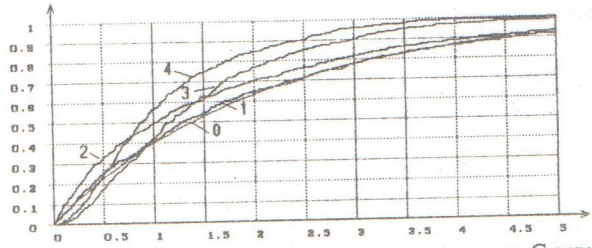

Рис.3. Эмпирические функции распределения статистики Смирнова при различном количестве оцениваемых параметров распределения Лапласа

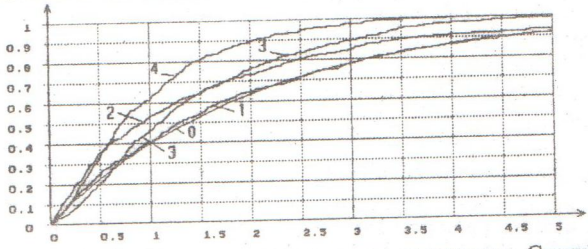

Рис.4. Эмпирические функции распределения статистики Смирнова при различном количестве оцениваемых параметров распределения Коши

Как видим, предельные распределения статистики Смирнова при условии оценивания параметров конкретного закона кардинально отличаются от распределения χ_2^2.

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 [6]. 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*. В табл. 1 сведены результаты идентификации предельных законов для статистики критерия Смирнова. Информация, представленная в таблице, должна интерпретироваться следующим образом. Указание в соответствующей клетке на конкретное распределение означает, что выборка соответствующей статистики хорошо описывается данным законом (согласуется с законом). Если в клетке таблицы содержится указание более, чем на один закон, то на первом месте стоит распределение, согласие с которым наилучшее. 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 таблице 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

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

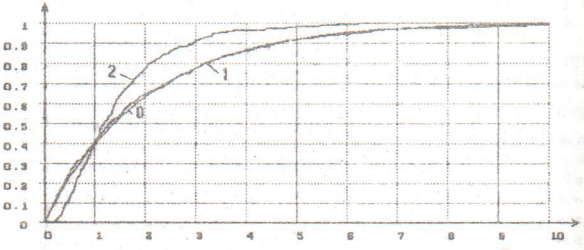

Рис.5. Эмпирические функции распределения статистики S_m Смирнова без оценивания параметров (*1*), при оценивании масштабного параметра (*2*) экспоненциального распределения

В данной работе исследовались и представлены в таблицах распределения статистик, когда наблюдаемые случайные величины распределены в соответствии с законами: нормальным с функцией плотности $f(x) = \frac{1}{\theta_1\sqrt{2\pi}} e^{-\frac{(x-\theta_0)^2}{2\theta_1^2}}$, Коши - $f(x) = \frac{\theta_0}{\pi[\theta_0^2 + (x - \theta_1)^2]}$, Лапласа - $f(x) = \frac{\theta_0}{2} e^{-\theta_0|x-\theta_1|}$, логистическим - $f(x) = \frac{\pi}{\theta_1\sqrt{3}} \exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\} \Big/ \left[1 + \exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2$, экспонен-

рмальным - $f(x)=\theta e^{-\theta x}$, полунормальным - $f(x)=\frac{2}{\theta\sqrt{2\pi}}e^{-x^2/2\theta^2}$, Рэлея - $f(x)=\frac{x}{\theta^2}e^{-x^2/2\theta^2}$.

Таблица 1.

Предельные распределения статистики Смирнова					
№ п/п	Распредел-е случайной величины	Параметры по выборке не оценивались	Оценивался только масштабный параметр	Оценивался только параметр сдвига	Оценивалось два параметра
1.	Нормальное	χ_2^2	γ(0.7737,0.4269,0.0024)	lnN(0.2761,0.5533) γ(1.9198,1.5578,0.3023)	lnN(0.1051,0.5478) γ(1.7555,1.7495,0.2892)
2.	Коши	χ_2^2	γ(0.7782,0.4814,0.0009)	γ(1.3746,0.9748,0.0213) lnN(-0.0058,0.9533)	γ(1.3257,1.3842,0.0149) lnN(-0.4064,0.9645)
3.	Лапласа	χ_2^2	γ(0.7744,0.4407,0.0021)	γ(1.4691,1.0715,0.0864) lnN(0.0930,0.7991)	lnN(-0.1539,0.8078) γ(1.8451,1.6025,0)
4.	Логистическое	χ_2^2	γ(0.8557,0.4890,0)	lnN(0.2770,0.4704) γ(2.5415,2.1924,0.3146)	lnN(0.0387,0.4431) γ(2.4369,2.8850,0.3043)
5.	Экспоненциальное	χ_2^2	lnN(0.1983,0.7328) γ(1.5151,1.0582,0.1554)		
6.	Полунормальное	χ_2^2	γ(1.2931,0.8505,0.1104) lnN(0.1502,0.8428)		
7.	Рэлея	χ_2^2	lnN(0.1983,0.7328) γ(1.5151,1.0582,0.1554)		

Полученные результаты моделирования подчеркивают, что 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 от распределений классических статистик, соответственно законов $K(s)$, χ_2^2, $a1(s)$ и $a2(s)$, что последние ни в коем случае не должны использоваться в такой ситуации.

На предельные распределения всех непараметрических статистик наиболее значительное влияние оказывает оценивание параметра сдвига, в существенно меньшей степени - оценивание масштабного параметра.

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

1. Орлов А.И. Распространенная ошибка при использовании критериев Колмогорова и омега-квадрат // Заводская лаборатория. 1985. Т. 51. № 1. С. 60-62.
2. Бондарев Б.В. О проверке сложных статистических гипотез // Заводская лаборатория. 1986. Т. 52. № 10. С. 62-63.
3. Durbin J. Kolmogorov-Smirnov tests when parameters are estimated // Lect. Notes Math. 1976. V.566. P.33-44.
4. Мартынов Г.В. Критерии омега-квадрат. - М.: Наука, 1978. - 80 с.
5. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - 416 с.
6. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. - Новосибирск: Изд-во НГТУ. - 1995. - 125 с.