УДК 519.24

А.И.Орлов, И.В.Орловский
(ЦЭМИ АН СССР)

ОЦЕНКА ОСТАТОЧНОГО ЧЛЕНА ПОРЯДКА n^{-2} ДЛЯ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ ДВУХВЫБОРОЧНОЙ СТАТИСТИКИ СМИРНОВА

Настоящая заметка является частью исследований по проблемам устойчивости в статистических процедурах и математических моделях реальных явлений. Здесь указано место рассматриваемой проблематики в общей схеме устойчивости.

В § I мы изучаем асимптотику распределения статистики $\mathcal{D}^{+}_{n,n}$, предложенной Н.В.Смирновым [I] для проверки совпадения функций распределения двух независимых выборок. В § 2 даются оценки остаточного члена порядка n^{-2}. В § 3 приводятся некоторые численные результаты. В § 4 мы обсуждаем место полученных результатов в прикладной математической статистике и обосновываем программу работ по получению аналогичных оценок для обычно используемых распределений.

§ I. Асимптотика распределения статистики $\mathcal{D}^{+}_{n,n}$

Рассмотрим две выборки объемов m и n из непрерывных функций распределения $\mathcal{F}(x)$ и $G(x)$ соответственно. Пусть $\mathcal{F}_m(x)$ и $G_n(x)$ - эмпирические функции распределения, построенные по этим выборкам. Для проверки гипотезы однородности выборок

$$H_0: \ \mathcal{F}(x) \equiv G(x) \tag{I}$$

Н.В.Смирнов в 1939 г. [I] предложил использовать статистику

$$\mathcal{D}^{+}_{m,n} = \sup\left(\mathcal{F}_m(x) - G_n(x)\right). \tag{2}$$

Мы будем рассматривать лишь случай равных объемов выборок: $m = n$. Относящийся к общему случаю обзор дан в работах [2], с. I3I-I35], [3].

Н.В.Смирнов показал, что

$$\lim_{n\to\infty} P\left\{\sqrt{n}\,\mathcal{D}^{+}_{n,n} > \lambda\right\} = e^{-\lambda^2}. \tag{3}$$

В 1951 г. Б.В.Гнеденко и В.С.Королюк [4] доказали, что для целого c

$$P\left\{\mathcal{D}^{+}_{n,n} \geqslant \frac{c}{n}\right\} = \frac{\binom{2n}{c+n}}{\binom{2n}{n}} . \tag{4}$$

Положим $\lambda = \frac{c}{\sqrt{n}}$ и введем в рассмотрение функцию

$$H(n,\lambda) = P\left\{\mathcal{D}^{+}_{n,n} \geqslant \frac{\lambda}{\sqrt{n}}\right\} . \tag{5}$$

Применив к правой части (4) формулу Стирлинга, можно получить асимптотическое разложение $H(n,\lambda)$ по степеням $n^{-1/2}$ при $n \to \infty$ [5]. Далее оцениваем погрешности асимптотических результатов при конечных n.

Нам понадобится следующий вариант формулы Стирлинга для гамма-функции $\Gamma(a)$ [6]:

$$\ln \Gamma(a) = \ln\sqrt{2\pi} + (a - \tfrac{1}{2})\ln a - a + \omega(a), \tag{6}$$

где

$$\omega(a) = \frac{1}{12a} - \frac{\theta}{360a^3} ,$$
$$0 < \theta = \theta(a) < 1 . \tag{7}$$

Выразим биномиальные коэффициенты в формуле (4) через гамма-функцию и применим формулу (6). После несложных преобразований получим

$$\ln H(n,\lambda) = -(n + \tfrac{1}{2})\ln(1 - \frac{\lambda^2}{n}) - \lambda\sqrt{n}\left(\ln(1 + \frac{\lambda}{\sqrt{n}}) - \right. \tag{8}$$
$$\left. - \ln(1 - \frac{\lambda}{\sqrt{n}})\right) + 2\omega(n) - \omega(n + \lambda\sqrt{n}) - \omega(n - \lambda\sqrt{n}) .$$

Для дальнейшего рассуждения нам понадобится ограничение на λ

$$0 \leq \frac{\lambda}{\sqrt{n}} \leq \gamma < 1 . \tag{9}$$

При справедливости ограничения (9) логарифмы в правой части равенства (8) можно разложить в ряды. Сделав это и приведя подобные члены, получим

$$H(n,\lambda)=\exp\{-\lambda^2\}\cdot\exp\left\{\frac{\varphi(\lambda)}{n}+\frac{R(n,\lambda)}{n^2}\right\}, \tag{10}$$

где
$$\varphi(\lambda)=\lambda^2\left(\frac{1}{2}-\frac{\lambda^2}{6}\right), \tag{11}$$

$$R(n,\lambda)=\lambda^4\sum_{k=2}^{\infty}\frac{\lambda^{2k-4}}{n^{k-2}}\left(\frac{1}{2k}-\frac{\lambda^2}{(2k+1)(k+1)}\right)- \tag{12}$$
$$-\frac{\lambda^2}{6\left(1-\frac{\lambda^2}{n}\right)}+\frac{Q}{360\,n},$$

$$Q=\frac{\theta(n+\lambda\sqrt{n})}{\left(1+\frac{\lambda}{\sqrt{n}}\right)^3}+\frac{\theta(n-\lambda\sqrt{n})}{\left(1-\frac{\lambda}{\sqrt{n}}\right)^3}-2\theta(n). \tag{13}$$

Воспользуемся разложением экспоненты в ряд Маклорена с остаточным членом в форме Лагранжа [7]. Получим далее

$$H(n,\lambda)=e^{-\lambda^2}\left(1+\frac{\varphi(\lambda)}{n}\right)+e^{-\lambda^2}\frac{\Delta}{n^2}, \tag{14}$$

где
$$\Delta=\Delta(n,\lambda)=\exp\left\{\theta_1\frac{\varphi(\lambda)}{n}\right\}\frac{\varphi^2(\lambda)}{2}+\exp\left\{\theta_2\frac{R(n,\lambda)}{n^2}+\right.$$
$$\left.+\frac{\varphi(\lambda)}{n}\right\}\cdot R(n,\lambda),\quad 0<\theta_i<1,\quad i=1,2. \tag{15}$$

Наша цель – получение равномерных по λ, n оценок Δ сверху и снизу в области, задаваемой (9) и условиями

$$0\le\lambda\le\lambda_{max};\qquad n\geqslant n_0. \tag{16}$$

Поскольку процентные точки λ_α предельного распределения для α = 0,1; 0,05; 0,01 равны соответственно 1,52; 1,73; 2,15, то для практических нужд достаточно рассматривать $\lambda_{max}\le 3$.

§ 2. Оценивание остаточного плана

Сначала рассмотрим $\varphi(\lambda)$ и $R(n,\lambda)$.

Легко убедиться, что $\varphi(\lambda)=0$ при $\lambda=0$ и $\lambda=\sqrt{3}$, положительна при $0<\lambda<\sqrt{3}$ и отрицательна при $\lambda>\sqrt{3}$, достигает максимума, равного $\frac{3}{8}$ при $\lambda=\sqrt{\frac{3}{2}}$, монотонно убывает при $\lambda>\sqrt{\frac{3}{2}}$. Поэтому при $\lambda_{max}>\sqrt{3}$

$$\max_{[0,\lambda_{max}]} \varphi(\lambda) = \frac{3}{8} \quad ; \quad \min_{[0,\lambda_{max}]} \varphi(\lambda) = \varphi(\lambda_{max}) . \tag{17}$$

Оценим $R(n,\lambda)$ сверху. Поскольку при $k \geqslant 2$

$$\frac{1}{2k} - \frac{\lambda^2}{(2k+1)(k+1)} \leq \frac{1}{2k} \leq \frac{1}{4} , \tag{18}$$

то сумма ряда (12) оценивается сверху геометрической прогрессией и

$$R(n,\lambda) \leq \frac{\lambda^4}{4\left(1-\frac{\lambda^2}{n}\right)} - \frac{\lambda^2}{6\left(1-\frac{\lambda^2}{n}\right)} + \frac{Q}{360\,n} . \tag{19}$$

Из ограничения (9) и равенства (13) следует, что

$$-2 < Q < 1 + \frac{1}{\left(1-\frac{\lambda}{\sqrt{n}}\right)^3} . \tag{20}$$

С помощью двух последних соотношений заключаем, что при $y \leq \frac{3}{4}$ и $n \geqslant 1$ справедлива оценка

$$R(n,\lambda) \leq \frac{1}{4\left(1-\frac{\lambda^2}{n}\right)} \left(\lambda^2 - \frac{1}{3}\right)^2 = R^{+}(n,\lambda) . \tag{21}$$

Оценим теперь

$$\max_{[0,\lambda_{max}]} R(n,\lambda) \cdot e^{\frac{\varphi(\lambda)}{n}} . \tag{22}$$

Используем при этом следующую лемму.

Л е м м а. Пусть $|b| \leq 1$; $0 < x_{max} < 1$; $1 \leq y_{max} \leq 9$. Функция

$$f(x,y) = \frac{(y+b)^2}{1-x}\, e^{x\left(\frac{1}{2}-\frac{1}{6}\right)} \tag{23}$$

в области $\{0 \leq x \leq x_{max}, 0 \leq y \leq y_{max}\}$ достигает своего максимума в точке $x = x_{max}$, $y = y_{max}$.

Доказательство основывается на вычислении частных производных. Нетрудно показать, что $\partial f/\partial x$ неотрицательна, а $\partial f/\partial y$ отрицательна при $y < -b$ и положительна при $y > -b$. Следовательно, максимум достигается или в точке $(x_{max}; 0)$ или в точке $(x_{max}; y_{max})$. Непосредственно сравнивая значения функции $f(x,y)$ в этих точках,

заключаем, что максимум достигается в точке $(x_{max}; y_{max})$

заключаем, что максимум достигается в точке $(x_{max}; y_{max})$.

Положим $x = \frac{\lambda^2}{n}$, $y = \lambda^2$, $b = -\frac{1}{3}$. Тогда $x_{max} = \gamma^2$, $y_{max} = \lambda^2$, $f(x,y) = 4R^+(n,\lambda)e^{\varphi(\lambda)/n}$. Из оценки (21) с помощью леммы получим

$$\max_{[0,\lambda_{max}]} R(n,\lambda)\exp\left\{\frac{\varphi(\lambda)}{n}\right\} \le \frac{1}{4}\,\frac{\left(\lambda_{max}^2 - \frac{1}{3}\right)^2}{1-\gamma^2}\, e^{\gamma^2\left(\frac{1}{2} - \frac{\lambda_{max}^2}{6}\right)} = \tag{24}$$
$$= h(\gamma, \lambda_{max}).$$

Теперь мы можем оценить Δ сверху, используя представление (15). Именно, из соотношений (15), (17), (21) и (24) следует, что

$$\Delta \le \exp\left\{\frac{3}{8n}\right\}\frac{\varphi^2(\lambda)}{2} + \exp\left\{\frac{R^+(n,\lambda)}{n^2}\right\} h(\gamma, \lambda_{max}). \tag{25}$$

При $\lambda > \sqrt{3}$ оценку (25) можно улучшить, поскольку в этом случае $\varphi(\lambda) < 0$ и

$$\Delta \le \frac{\varphi^2(\lambda)}{2} + \exp\left\{\frac{R^+(n,\lambda)}{n^2}\right\}\cdot h(\gamma, \lambda_{max}). \tag{26}$$

Оценим $R(n,\lambda)$ снизу. Пусть

$$a_k(\lambda) = \frac{1}{2k} - \frac{\lambda^2}{(2k+1)(k+1)}; \quad k = 2, 3, \ldots \tag{27}$$

Нетрудно показать, что $a_k(\lambda) \geqslant a_2(\lambda)$ при $\lambda^2 \geqslant 3{,}75$ и $a_k(\lambda) \geqslant 0$ при $\lambda^2 \le 3{,}75$. Из соотношений (12), (20) получаем оценку

$$R(n,\lambda) \geqslant \frac{1}{\left(1 - \frac{\lambda^2}{n}\right)}\left(\lambda^4\left(\frac{1}{4} - \frac{\lambda^2}{15}\right) - \frac{\lambda^2}{6}\right) - \frac{1}{180} = R^-(n,\lambda) \tag{28}$$

при $\lambda^2 \geqslant 3{,}75$, т.е. при $\lambda \geqslant 1{,}94$, и

$$R(n,\lambda) \geqslant -\frac{\lambda^2}{6\left(1 - \frac{\lambda^2}{n}\right)} - \frac{1}{180} = R^-(n,\lambda) \tag{29}$$

при $\lambda \le 1{,}94$. Отметим, что $\Delta < 0$ только в случае $R(n,\lambda) < 0$; а тогда $\exp\left\{\theta_2 \frac{R(n,\lambda)}{n^2}\right\} < 1$. С помощью этого замечания и соотношений (15), (28), (29) заключаем, что при $\lambda > \sqrt{3}$

$$\Delta \geqslant exp\left\{\frac{\varphi(\lambda)}{n}\right\}\frac{\varphi^2(\lambda)}{2} + R^-(n,\lambda). \qquad (30)$$

При $\lambda \leq \sqrt{3}$ оценка выглядит несколько иначе:

$$\Delta \geqslant \frac{\varphi^2(\lambda)}{2} + exp\left\{\frac{\varphi(\lambda)}{n}\right\}R^-(n,\lambda). \qquad (31)$$

Из неравенств (25), (26), (30), (31) нетрудно получить равномерные оценки по области (9), (16). При этом естественно рассматривать интервалы $[0;\sqrt{3})$ и $[\sqrt{3};\lambda_{max}]$ отдельно. Поскольку отклонения быстро растут с увеличением λ_{max}, рассмотрим только оценивание Δ на втором интервале.

Из (21) следует, что при $\sqrt{3} \leq \lambda \leq \lambda_{max}$

$$R^+(n,\lambda) \leq \frac{1}{4(1-\gamma^2)}\left(\lambda_{max}^2 - \frac{1}{3}\right)^2 = g(\lambda_{max};\gamma). \qquad (32)$$

Из (16), (26) и (32) следует, что при всех $n \geqslant n_0$ и $\sqrt{3} \leq \lambda \leq \lambda_{max}$ справедлива оценка

$$\Delta(n,\lambda) \leq \frac{\varphi^2(\lambda_{max})}{2} + exp\left\{\frac{g(\lambda_{max},\gamma)}{n_0^2}\right\}\cdot h(\gamma,\lambda_{max}). \qquad (33)$$

Нижняя граница находилась при численной минимизации по λ, полученной из неравенства (30) следующей нижней оценки для $\Delta(n,\lambda)$:

$$p(\lambda) = exp\left\{\frac{\varphi(\lambda_{max})}{n_0}\right\}\frac{\varphi^2(\lambda)}{2} + R_\gamma^-(n,\lambda). \qquad (34)$$

Здесь $R_\gamma^-(n,\lambda)$ отличается от $R^-(n,\lambda)$ лишь заменой $1-\frac{\lambda^2}{n}$ в знаменателе (см. неравенства (28), (29)) на $1-\gamma^2$.

§ 3. Численные результаты

При нескольких значениях λ_{max}, γ, n_0 изложенными в § 2 методами были вычислены Δ_1 и Δ_2 такие, что

$$\Delta_1 < \Delta(n,\lambda) < \Delta_2 \qquad (35)$$

для всех λ, n из области, задаваемой неравенствами (9), (16).

Результаты сведены в таблицу. Напомним, что $\lambda\sqrt{n}$ во всех рассуждениях предполагается целым.

	λ_{max}	γ	n_0	Δ_1	Δ_2
I	3	0,5	8	- I6,5	69,6
2	2,5	0,75	8	- 6,2	26,0
3	I,73	0,5	8	- 0,7I	2,65
4*)	3	0,5	I	- 86	I2I

Рассмотрим пример применения полученных оценок остаточного члена. Для данных [9], относящихся к изучению влияния консервационной среды на момент трения подшипника, n = I0, c = 5, $D^{+}_{10,10}$ = 0,5, $\lambda = c/\sqrt{n}$ = I,58. С помощью равенства (4) можно подсчитать, что $P(D^{+}_{10,10} \geqslant 0,5)$ = 0,0839, в то время как в соответствии с равенством (3) предельное значение $e^{-\lambda^2} = e^{-2,5}$ = 0,082I. Наши результаты (формула (I4) и третья строка таблицы) позволяют заключить, что интересующая нас вероятность лежит между 0,083I и 0,0858. По нашему мнению, достигнутую точность следует считать удовлетворительной, и изучать члены порядка n^{-3} с целью применения в обработке реальных данных нет необходимости.

Покажем на этом же примере, что широта границ для Δ связана в основном с использованием равномерных оценок, а не с малым числом оставленных членов в асимптотическом разложении. Для этого рассмотрим равенство (I5) при n = I0, λ = I,58. Тогда $\varphi(\lambda)$ = 0,208, $R(n,\lambda)$ = I,43. Полагая (I5) $\theta_1 = \theta_2$ = = I, находим верхнюю границу для Δ, равную I,502. Полагая $\theta_1 = \theta_2$ = 0, находим нижнюю границу Δ, равную I,482. Значит, погрешность в определении Δ, вызванная преобразованиями, приведшими в появлению θ_1 и θ_2, есть I,502 - I,482 = = 0,020, в то время как погрешность, порожденная использованием равномерных оценок, равна 2,65 - (- 0,7I) = 3,36, т.е. в I68 раз больше.

§ 4. Перспективы

Пусть $S_n(z)$ - функция распределения некоторой статистики, построенной по выборке объема n. Часто $S_n(z)$ неизвестна

*) Результаты строки 4 были получены более грубыми методами [8], чем изложенные в § 2 настоящей заметки.

или выражена весьма громоздко. Тогда при обработке реальных данных рекомендуют пользоваться предельным распределением $S_0(z)$ или (реже) отрезком асимптотического разложения. Подобная рекомендация неявно предполагает, что соответствующие остаточные члены малы. Однако это предположение требует доказательства. По нашему мнению, для применяемых обычно статистик необходимо иметь удобные оценки остаточных членов. Нам представляется, что полученные в настоящей заметке оценки являются достаточно практичными.

Отметим несколько работ, наличие которых стимулировало появление настоящей заметки. С.Н.Бернштейн в ряде статей [10], [11] и В.Феллер [12] получали оценки биномиального распределения с помощью нормального. Ряд исследователей оценивал константу в интервале Берри-Эссеена [13, с. 621]. Лучшие результаты получены В.М.Золотаревым [14].

Если известны точные формулы для $S_n(z)$, что имеет место для $D^+_{m,n}$ при $m = np$, p – целое ([2], [15], с. 185), одновыборочной статистики Смирнова [16], распределений биномиального, Пуассона, мультиномиального, отрицательного биномиального, то для получения оценок, аналогичных приведенным в настоящей заметке, необходимо проделать простые, но весьма кропотливые вычисления, связанные с применением формулы Стирлинга, оценками остаточных членов в рядах, сглаживанием функции распределения с помощью асимптотического разложения типа Эйлера-Маклорена. Могут оказаться полезными развитые В.М.Калининым [17], [18] методы получения асимптотических разложений. В частности, было бы интересно получить в асимптотическом разложении биномиального распределения равномерную оценку остаточного члена порядка $n^{-3/2}$, т.е. по порядку лучшую, чем оценки Бернштейна-Феллера и Дж.Успенского [19], с. 129.

Однако часто оказывается затруднительным даже оценить скорость убывания

$$\Delta_n = \sup_z |S_n(z) - S_0(z)| .$$

Так, в случае статистики χ^2 мы приходим к известной теоретико-числовой задаче о числе целых точек в эллипсоиде, которая к настоящему времени еще не решена окончательно [20], [21], [22]. Для статистик типа ω^2 доказано лишь, что $\Delta_n = O(n^{-1/2} \ln n)$, хотя есть основания полагать $\Delta_n = O(n^{-1})$ [23], [24].

Для оценки остаточных членов естественно использовать ЭВМ.

Однако для обоснования применимости результатов счета желательно доказать монотонность $\Delta(n,\lambda)$ по n при фиксированном λ.

Мы считаем необходимым развернуть исследования по получению удобных оценок остаточных членов в асимптотических разложениях обычно применяемых распределений. В силу отмеченной выше идейной простоты вычислений многое может быть сделано в рамках дипломных и курсовых работ.

Литература

1. Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках. - Бюлл. МГУ, сер. А, 2, 1939, вып. 2, 3-14.

2. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М., Наука, 1965.

3. Гаек Я., Шидак З. Теория ранговых критериев. М., Наука, 1971.

4. Гнеденко Б.В., Королюк В.С. О максимальном расхождении двух эмпирических распределений. - ДАН СССР, 1951, 80, 4, 525-528.

5. Боровков А.А. К задаче о двух выборках. - Изв. АН СССР, сер. матем., 1962, 26, 605-624.

6. Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. М., Наука, 1966 Т. II.

7. Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. М., Наука, 1966 Т. I.

8. Орлов А.И. Оценка остаточного члена для функции распределения двухвыборочной статистики Смирнова. - В сб. Алгоритмы многомерного статистического анализа и их применения. М., ЦЭМИ АН СССР, 1975, 105-108.

9. Егорова Л.А., Харитонов Ю.С., Соколовская Л.В. О применении непараметрического χ -критерия ван-дер-Вардена при статистической обработке результатов наблюдений. - Заводская лаборатория, 1976, 42, 10, 1237-1237.

10. Бернштейн С.Н. Об одном видоизменении неравенства Чебышева и о погрешности формулы Лапласа. - Учен. зап. научно-исслед. кафедр Украины, отд. матем., 1924, вып. I, 38-48.

11. Бернштейн С.Н. Возврат к вопросу о точности предельной формулы Лапласа. - Изв. АН СССР, сер. матем., 1943, 7, 3-14.

12. Feller W. On the Normal Approximation to the Binomial Distribution.- Ann.Math.Stat., 16, 4, 1945, 319-329.

13. Ф е л л е р В. Введение в теорию вероятностей и ее приложения. М., Мир, 1967, Т. 2.

14. З о л о т а р е в В.М. Некоторые неравенства теории вероятностей и их применения к уточнению теоремы А.М.Ляпунова. - ДАН СССР, 1967, 177, 3, 501-504.

15. Т а к а ч Л. Комбинаторные методы в теории случайных процессов. М., Мир, 1971.

16. С м и р н о в Н.В. Приближение законов распределения случайных величин по эмпирическим данным. - Усп. математ. наук, 1944, 10, 179-206.

17. К а л и н и н В.М. Предельные свойства вероятностных распределений. - Труды Матем. ин-та им. В.А.Стеклова. Л., Наука, 1968. Т. 104, 88-134.

18. К а л и н и н В.М. Гамма-функция и вероятностные предельные теоремы. - Труды Матем. ин-та им. В.А.Стеклова. Л., Наука, 1970. Т. Ш, 163-194.

19. Uspensky I. Introduction to Mathematical Probability. N.Y., 1937.

20. Esseen C.G. Fourier Analysis of Distribution Functions. A Mathematical Study of the Laplace-Gaussian Law. Acta Math., 77, 1945, 1-125.

21. К а л и н и н В.М., Ш а л а е в с к и й О.В. Исследования по классическим проблемам теории вероятностей и математической статистики. П. - Записки научн. сем. Ленингр. отд. Матем. ин-та им. В.А.Стеклова. Л., Наука, 1972, Т. 26.

22. О р л о в А.И. Предельные теоремы для статистик интегрального типа. - Международная конференция по теории вероятностей и математической статистике (Вильнюс, 25-30 июня, 1973). Тезисы докладов. Вильнюс, 1973. Т. 2, 137- 140.

23. О р л о в А.И. Скорость сходимости распределения статистики Мизеса-Смирнова. - Теория вероятн. и ее примен., XIX, 4, 1974, 766-786.

24. Csorgo S. On an asymptotic expansion for the Von Mises w^2 statistic. Acta Sci.Math., 38, 1-2, 1976, 45-67.