УДК 519.2

В. И. Тимонин

АНАЛОГИ ДВУХВЫБОРОЧНЫХ СТАТИСТИК РЕНЬИ ДЛЯ ПРОВЕРКИ ГИПОТЕЗЫ ЛЕМАНА

Рассмотрена двухвыборочная задача проверки степенной гипотезы Лемана для цензурированных справа выборок. Предложен непараметрический критерий проверки этой гипотезы, являющийся аналогом критерия Реньи однородности двух выборок. Получены точные и предельные распределения статистики критерия при справедливости рассматриваемых гипотез.

В работе [1] была рассмотрена следующая задача. Пусть имеются две независимые выборки $\bar{x} = (x_1, \ldots, x_m)$, $\bar{y} = (y_1, \ldots, y_n)$, причем $x_i \sim F(x)$, $y_j \sim G(x)$, $i = \overline{1,m}$, $j = \overline{1,n}$. Требуется проверить основную (нулевую) гипотезу

$$H_0\colon F(x) = (G(x))^k, \tag{1}$$

где $k \geqslant 1$ — известное фиксированное число.

Гипотезы вида (1) рассматривались впервые Леманом в работе [2] в качестве альтернативных к гипотезе однородности. Позднее Кокс [3] исследовал методы оценки зависимостей числа k от факторов (ковариат), предполагая, что выполнена гипотеза (1) (точнее, аналогичное соотношение для функций $\bar{F}(x) = 1 - F(x), \bar{G}(x) = 1 - G(x)$).

В работе [1] был предложен критерий проверки гипотезы (1), а также получены предельное распределение его статистики при выполнении гипотезы (1) и метод вычисления ее точных распределений. В настоящей работе исследуется задача проверки гипотезы (1) в том случае, когда выборки $\bar{x}$, $\bar{y}$ являются цензурированными справа. Цензурирование часто имеет место при испытаниях технических систем, в клинических исследованиях и т.д.

При проверке однородности (т.е. при $k = 1$) в таких случаях наиболее часто применяют двухвыборочный критерий Реньи, статистика которого имеет вид [4]

$$R_q = \sqrt{\frac{(1-q)mn}{q(m+n)}} \max_{\hat{H}_{m+n} < q} \frac{|\hat{F}_m(x) - \hat{G}_n(x)|}{1 - \hat{H}_{m+n}}, \tag{2}$$

где $\hat{F}_m(x), \hat{G}_n(x)$ — эмпирические функции распределения выборок $\bar{x}$, $\bar{y}$;

$$\hat{H}_{m+n}(x) = \frac{1}{m+n}\left(m\hat{F}_m(x) + n\hat{G}_n(x)\right)$$

— объединенная эмпирическая функция распределения; q — фиксированное число, $0 < q < 1$.

Реньи доказал, что предельное распределение R_q при условии $k = 1$ не зависит от параметра q и имеет вид

$$L(x) = \frac{4}{\pi}\sum_{i=0}^{\infty}\frac{(-1)^i}{2i+1}\exp\left(-\frac{(2i+1)^2\pi^2}{8x^2}\right). \qquad (3)$$

Далее после введения необходимых обозначений определим статистику R_{qk} и докажем, что ее предельное распределение не зависит от параметров q и k. Затем рассмотрим метод вычисления ее точных распределений для произвольных q, k. В заключение приведем таблицы точных значений вероятностей $P(T_{qk} < h)$ для ряда значений m, n, q, k, h и изучим вопрос допустимости использования предельного распределения при небольших объемах выборок $\bar{x}$, $\bar{y}$.

Без ограничения общности будем считать, что $F(x)=x$, $G(x)=x^{k_0}$, $k_0 = 1/k$, $0 \leqslant x \leqslant 1$. Аналогом функции $\hat{H}_{m+n}(x)$ является функция

$$h_{mn} = \frac{1}{m+n}\left(m\hat{F}_m(x) + n(\hat{G}_n(x))^k\right).$$

Далее для упрощения записи будем часто опускать индексы m, n функции $h_{mn}(x)$, обозначая ее через $h(x)$.

Для дальнейшего изложения потребуется следующее утверждение.

Лемма. *Пусть выполнена гипотеза* (1). *Тогда при* $m, n \to \infty$ *справедливо равенство*

$$P(\sup_x |h_{mn}(x) - x| \to 0) = 1.$$

Доказательство очевидно в силу теоремы Гливенко и непрерывности на отрезке [0, 1] функции $y = x^k$.

В работе [1] доказана теорема, которая используется при выводе основного результата.

Теорема 1. *При* $m, n \to \infty$, $m/n \to \rho > 0$ *распределение случайного процесса*

$$Z_{mn}(x) = \sqrt{m}(\hat{F}_m(x) - (\hat{G}_n(x))^k), \quad 0 \leqslant x \leqslant 1,$$

слабо сходится к распределению непрерывного гауссовского процесса $Z(x)$ *с характеристиками*

$$EZ(x) = 0, \quad EZ(s)Z(t) = s(1-t+\mu t^{1-k_0}(1-t^{k_0})), \quad \mu = k^2\rho, \quad s \leqslant t. \qquad (4)$$

Рассмотрим статистику

$$T_{qk} = \sqrt{\frac{(1-q)m}{q}} \max_{\varphi(h)<q} \frac{|\hat{F}_m(x) - (\hat{G}_n(x))^k|}{1 - h(x) + \mu((h(x))^{1-k_0} - h(x))}, \qquad (5)$$

где

$$\varphi(h) = \frac{h}{1 + \mu(h^{1-k_0} - h)}, \quad 0 < q < 1.$$

Для предельного распределения статистики (5) справедлива следующая теорема.

Теорема 2. *При $m, n \to \infty$, $m/n \to \rho > 0$ предельное распределение статистики T_{qk} не зависит от параметров q, k и совпадает с распределением Реньи (3).*

Доказательство. Заметим прежде всего, что из условия $\varphi(h) < q$ следует, что знаменатель дроби в статистике (5) ограничен снизу:

$$1 - h(x) + \mu((h(x))^{1-k_0} - h(x)) = \\ = (1 - \varphi(h(x)))(1 + \mu((h(x))^{1-k_0} - h(x))) \geqslant 1 - q > 0.$$

В этом случае, как показано в работах [4, 5], в силу леммы при выводе асимптотического распределения можно заменить $h_{mn}(x)$ на предельную функцию x, $0 \leqslant x \leqslant 1$. С учетом теоремы 1 получим, что асимптотическое распределение статистики (5) совпадает с распределением функционала вида

$$\Phi(Z(x)) = \sqrt{\frac{1-q}{q}} \sup_{\varphi(x)<q} \frac{|Z(x)|}{1 - x + \mu(x^{1-k_0} - x)}.$$

Рассмотрим строго возрастающее преобразование

$$y = \varphi(x) = \frac{x}{1 + \mu(x^{1-k_0} - x)}, \quad [0,\ 1] \to [0,\ 1].$$

Пусть $x(y)$ — обратное преобразование. Тогда процесс

$$W(y) = \frac{Z(x(y))}{1 + \mu(x(y))^{1-k_0} - x(y)}$$

является броуновским мостом, что следует из равенств

$$EW(y) = 0, \quad EW(u)W(v) = \frac{x(u)}{1 + \mu((x(u))^{1-k_0} - x(u))} \times \\ \times \left(\frac{1 - x(v) + \mu((x(v))^{1-k_0} - x(v))}{1 + \mu((x(v))^{1-k_0} - x(v))} \right) = u(1-v), \quad 0 \leqslant u \leqslant v \leqslant 1.$$

Учитывая, что распределение функционалов типа экстремумов процессов не изменяется при монотонном изменении шкалы времени, получим при $x = x(y)$

$$P\left(\sqrt{\frac{1-q}{q}}\sup_{\varphi(x)<q}\frac{|Z(x)|}{1-x+\mu(x^{1-k_0}-x)}<h\right)=$$

$$=P\left(\sqrt{\frac{1-q}{q}}\sup_{y<q}|W(y)|\frac{1+\mu((x(y))^{1-k_0}-x(y))}{1-x(y)+\mu((x(y))^{1-k_0}-x(y))}<h\right)=$$

$$=P\left(\sqrt{\frac{1-q}{q}}\sup_{y<q}\frac{|W(y)|}{1-\dfrac{x(y)}{1+\mu((x(y))^{1-k_0}-x(y))}}<h\right)=$$

$$=P\left(\sqrt{\frac{1-q}{q}}\sup_{y<q}\frac{|W(y)|}{1-y}<h\right).$$

Последняя вероятность является предельным значением вероятности $P(R_q < h)$ для статистики Реньи (2).

Доказанная теорема позволяет проверить гипотезу (1) при больших объемах выборок m, n. На практике, особенно при испытаниях технических систем, количество образцов никогда не превышает нескольких десятков. Как будет показано далее, в этом случае использование предельного распределения может привести к значительным ошибкам при проверке гипотезы (1). По этим причинам важное значение имеет задача вычисления точных распределений статистики (5).

В работе [6] приведен метод вычисления точных распределений — статистик типа Колмогорова–Смирнова для случая l выборок, функции распределения которых связаны степенной зависимостью. Он основан на теории случайного блуждания по ячейкам l-мерной матрицы A; значения функции распределения статистик равны вероятности невыхода траекторий блуждания за пределы некоторого подмножества $A_0 \subset A$. Применение этого метода к рассматриваемой статистике составляет содержание приводимой далее теоремы. Поскольку ее утверждение является простым следствием более общего результата из работы [6], то доказательство опускается.

Теорема 3. *Вероятность $P(T_{qk} < h)$ равна величине $\pi_{mn}(h)$, которая может быть вычислена с помощью итерационной процедуры*

$$\pi_{ij}(h)=\left(\pi_{i-1,j}\frac{ik}{ik+j}+\pi_{i,j-1}\frac{j}{ik+j}\right)\chi_{ij}(A_0) \qquad (6)$$

с начальными и граничными условиями

$$\pi_{00}(h)=1,\quad \pi_{-1,i}(h)=0,\quad \pi_{i,-1}(h)=0,\quad i=\overline{0,m},\quad j=\overline{0,n};$$

здесь

$$h > 0, \quad \chi_{ij} = \begin{cases} 1, & если \ \ (i,j) \in A_0, \\ 0, & если \ \ (i,j) \notin A_0. \end{cases}$$

Множество A_0 состоит из пар целых чисел (i,j), $i = \overline{0,m}$, $j = \overline{0,n}$, для которых справедливо хотя бы одно из следующих условий:

а) $$\frac{a_{ij}}{1+\mu(a_{ij}^{1-k_0} - a_{ij})} \geqslant q, \quad a_{ij} = \frac{in^{k-1} + j^k}{(m+n)n^{k-1}};$$

б) $$\left\{ \frac{a_{ij}}{1+\mu(a_{ij}^{1-k_0} - a_{ij})} < q \right\} \cap$$

$$\cap \left\{ \sqrt{\frac{1-q}{q}} \sqrt{m} \frac{\left| \frac{i}{m} - \left(\frac{j}{n} \right)^k \right|}{1 - a_{ij} + \mu \left(a_{ij}^{1-k_0} - a_{ij} \right)} < h \right\}.$$

Метод позволяет рассчитывать точные распределения для больших объемов выборок m, n ввиду отсутствия в процедуре (6) больших или малых множителей.

В табл. 1, 2 приведены рассчитанные точные значения вероятностей $P(T_{qk} < h)$ для двух значений глубины цензурирования q: $q_1 = 0{,}7$; $q_2 = 0{,}85$. Три значения аргумента $h_1 = 1{,}78$; $h_2 = 1{,}96$; $h_3 = 2{,}24$ выбраны как наиболее близкие к квантилям уровня 0,85; 0,9; 0,95 предельного распределения Реньи (3). Объемы выборок полагаются равными $m = n$ и изменяются в пределах $20 \leqslant m \leqslant 10000$. В таблицах приведены первые четыре цифры после запятой, т.е. если вероятность равна 0,83246, то в таблице записано 0,8325.

Анализ результатов расчета показывает очень медленную сходимость вероятностей к их предельным значениям. Даже при объемах выборок порядка 1000 разница может превышать 0,01. При объемах от 20 до 50 разница нередко составляет величину, доходящую до 0,1, причем точные вероятности всегда больше предельных. На практике, когда используются в основном асимптотические результаты, это приводит к существенному увеличению вероятности ошибок первого рода.

Сформулированные выводы справедливы не только для статистик (5). Оказалось, что медленная сходимость к предельным вероятностям имеет место и для $k = 1$ при использовании статистик (2). Применяя несколько измененную процедуру (6), были вычислены точные вероятности $P(R_q < 1{,}78)$ при $m = n = 5000$ и $m = n = 10000$, когда $q = 0{,}9$.

Таблица 1

Значения вероятностей $P(T_{kq} < h)$ при $q = 0{,}7$

m	При $k = 1{,}5$			При $k = 2$			При $k = 3$		
	$h =$ $= 1{,}78$	$h =$ $= 1{,}96$	$h =$ $= 2{,}24$	$h =$ $= 1{,}78$	$h =$ $= 1{,}96$	$h =$ $= 2{,}24$	$h =$ $= 1{,}78$	$h =$ $= 1{,}96$	$h =$ $= 2{,}24$
20	0,9290	0,9589	0,9599	0,9013	0,9014	1	1	1	1
30	0,8978	0,9295	0,9762	0,9198	0,9569	0,9696	0,8308	0,9992	1
40	0,8836	0,9407	0,9652	0,8734	0,9396	0,9754	0,8945	0,9029	0,9998
50	0,8861	0,9312	0,9593	0,8983	0,9321	0,9668	0,9319	0,9408	0,9527
60	0,8802	0,9265	0,9618	0,8919	0,9427	0,9755	0,9183	0,9522	0,9725
70	0,8858	0,9220	0,9633	0,8952	0,9259	0,9666	0,9086	0,9542	0,9817
80	0,8835	0,9207	0,9652	0,8934	0,9398	0,9691	0,8758	0,9331	0,9736
90	0,8808	0,9249	0,9624	0,8990	0,9256	0,9707	0,8846	0,9255	0,9661
100	0,8723	0,9221	0,9640	0,8799	0,9270	0,9673	0,8956	0,9349	0,9593
200	0,8715	0,9113	0,9567	0,8748	0,9099	0,9601	0,8794	0,9192	0,9646
300	0,8650	0,9135	0,9560	0,8671	0,9187	0,9559	0,8645	0,9201	0,9596
400	0,8633	0,9088	0,9557	0,8617	0,9119	0,9570	0,8651	0,9149	0,9551
500	0,8623	0,9084	0,9550	0,8649	0,9096	0,9547	0,8668	0,9109	0,9557
600	0,8619	0,9092	0,9559	0,8613	0,9098	0,9558	0,8644	0,9120	0,9563
700	0,8607	0,9073	0,9560	0,8591	0,9091	0,9547	0,8647	0,9112	0,9570
800	0,8605	0,9074	0,9536	0,8601	0,9080	0,9546	0,8616	0,9114	0,9566
900	0,8592	0,9071	0,9534	0,8590	0,9089	0,9540	0,8634	0,9119	0,9562
1000	0,8572	0,9061	0,9531	0,8595	0,9076	0,9549	0,8624	0,9070	0,9543
2000	0,8561	0,9044	0,9521	0,8571	0,9051	0,9527	0,8564	0,9058	0,9539
3000	0,8545	0,9033	0,9516	0,8546	0,9041	0,9519	0,8572	0,9053	0,9526
5000	0,8535	0,9029	0,9514	0,8541	0,9034	0,9517	0,8550	0,9040	0,9520
7500	0,8527	0,9023	0,9511	0,8532	0,9025	0,9512	0,8541	0,9032	0,9514
10000	0,8525	0,9019	0,9509	0,8526	0,9022	0,9510	0,8534	0,9027	0,9513
∞	0,8498	0,9000	0,9498	0,8498	0,9000	0,9498	0,8498	0,9000	0,9498

Таблица 2

Значения вероятностей $P(T_{kq} < h)$ при $q = 0{,}85$

m	При $k = 1{,}5$			При $k = 2$			При $k = 3$		
	$h =$ $= 1{,}78$	$h =$ $= 1{,}96$	$h =$ $= 2{,}24$	$h =$ $= 1{,}78$	$h =$ $= 1{,}96$	$h =$ $= 2{,}24$	$h =$ $= 1{,}78$	$h =$ $= 1{,}96$	$h =$ $= 2{,}24$
20	1	1	1	1	1	1	1	1	1
30	0,8209	1	1	1	1	1	1	1	1
40	0,9355	0,9370	1	0,8443	1	1	1	1	1
50	0,9260	0,9630	0,9676	0,8855	0,8855	1	1	1	1
60	0,8965	0,9391	0,9819	0,9258	0,9367	0,9999	1	1	1
70	0,8821	0,9384	0,9747	0,9156	0,9579	0,9624	0,8616	1	1
80	0,9012	0,9450	0,9664	0,9327	0,9578	0,9739	0,8602	0,9998	1
90	0,9065	0,9438	0,9717	0,9045	0,9557	0,9841	0,8920	0,8925	1
100	0,8781	0,9340	0,9733	0,8810	0,9336	0,9845	0,9076	0,9164	0,9999
200	0,8842	0,9212	0,9632	0,8874	0,9203	0,9720	0,8977	0,9123	0,9823
300	0,8727	0,9171	0,9583	0,8901	0,9220	0,9632	0,8809	0,9108	0,9733
400	0,8730	0,9188	0,9591	0,8766	0,9163	0,9627	0,8867	0,9244	0,9621
500	0,8679	0,9123	0,9573	0,8736	0,9162	0,9603	0,8847	0,9215	0,9641
600	0,8636	0,9122	0,9578	0,8706	0,9144	0,9577	0,8756	0,9254	0,9591
700	0,8611	0,9135	0,9566	0,8636	0,9152	0,9593	0,8741	0,9202	0,9583
800	0,8662	0,9139	0,9569	0,8686	0,9128	0,9577	0,8726	0,9165	0,9609
900	0,8640	0,9105	0,9556	0,8664	0,9118	0,9567	0,8698	0,9124	0,9595
1000	0,8609	0,9086	0,9550	0,8664	0,9088	0,9558	0,8735	0,9124	0,9600
2000	0,8582	0,9070	0,9531	0,8607	0,9075	0,9543	0,8637	0,9106	0,9543
3000	0,8570	0,9050	0,9528	0,8590	0,9073	0,9525	0,8628	0,9055	0,9541
5000	0,8550	0,9043	0,9522	0,8565	0,9044	0,9524	0,8576	0,9054	0,9539
7500	0,8538	0,9032	0,9517	0,8556	0,9036	0,9522	0,8569	0,9053	0,9524
10000	0,8536	0,9027	0,9515	0,8541	0,9033	0,9518	0,8553	0,9041	0,9519
∞	0,8498	0,9000	0,9498	0,8498	0,9000	0,9498	0,8498	0,9500	0,9498

Предельное значение вероятности равно 0,8498, точные — 0,8532 ($n = 10000$). Разница превышает 0,003.

В заключение автор выражает признательность И.И. Барышниковой за составление программы, реализующей алгоритм (6).

СПИСОК ЛИТЕРАТУРЫ

1. Тимонин В. И. О предельном распределении статистики одного непараметрического критерия // Теория вероятностей и ее применение. – 1987. – Т. 32. – № 4. – С. 790–792.
2. Lehman E. The Power of Rank Tests // Annals of Mathematical Statistics. – 1953. – V. 24. – № 1. – P. 23–43.
3. Cox D. Regression Models and Life-Tables // J. Royal Statist. Society. – 1972. – V. B-34. – P. 187–220.
4. Гаек Я., Шидак З. Теория ранговых критериев. – М.: Наука. – 1971. – 376 с.
5. Королюк В. С., Боровских Ю. В. Аналитические проблемы асимптотики вероятностных распределений. – Киев: Наукова думка, 1981. – 240 с.
6. Тимонин В. И., Черномордик О. М. Метод вычисления точного распределения статистик типа Колмогорова–Смирнова при альтернативах Лемана // Теория вероятностей и ее применение. – 1985. – Т. 30. – № 3. – С. 572–573.

Статья поступила в редакцию 4.06.2004

Владимир Иванович Тимонин родился в 1952 г., окончил в 1975 г. Московский институт электронного машиностроения. Канд. физ.-мат. наук, доцент кафедры “Высшая математика" МГТУ им. Н.Э. Баумана. Автор 26 научных работ в области теории надежности и математической статистики.

V.I. Timonin (b. 1952) graduated from the Moscow Institute for Electronic Engineering in 1975. Ph.D. (Phys.-Math.), assoc. professor of “Higher Mathematics” department of the Bauman Moscow State Technical University. Author of 26 publications in the field of theory of reliability and mathematical statistics.