

Модифицированные критерии согласия Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга для случайно цензурированных выборок. Ч. 1*

Б.Ю. ЛЕМЕШКО, Е.В. ЧИМИТОВА, М.А. ВЕДЕРНИКОВА

Методами компьютерного моделирования исследуются распределения статистик модифицированных критериев Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга при различных объемах выборок и распределениях моментов цензурирования. Рассматриваются случаи проверки простых и сложных гипотез о согласии. Формулируется алгоритм моделирования случайно цензурированных выборок в случае неизвестного распределения моментов цензурирования. Описывается пример проверки сложной гипотезы о согласии по случайно цензурированной выборке с использованием рассматриваемых критериев.

Ключевые слова: случайное цензурирование, модифицированные критерии согласия типа Колмогорова, Крамера–Мизеса–Смирнова, Андерсона–Дарлинга, оценка Каплана–Мейера.

ВВЕДЕНИЕ

Важнейшим этапом при построении вероятностной модели является проверка принадлежности наблюдаемой случайной величины предполагаемому закону распределения вероятностей. Проверка такого рода гипотез осуществляется с использованием критериев согласия. Положительный результат проверки позволяет утверждать, что использование построенной модели в дальнейшем не приведет к существенным ошибкам.

Проверяемые гипотезы могут быть простыми и сложными. *Простая* проверяемая гипотеза имеет вид $H_0: F(x) = F_0(x; \theta)$, где значения параметров $\theta = (\theta_1, \theta_2, \dots, \theta_m)^T$ закона $F_0(x; \theta)$ известны. *Сложная* гипотеза имеет вид $H_0: F(x) \in \{F_0(x; \theta), \theta \in \Omega_\theta\}$. Применение критерия согласия в случае проверки сложной гипотезы отличается от случая проверки простой гипотезы, если оценки неизвестных параметров вычисляются по той же выборке, по которой проверяется гипотеза о согласии.

В задачах анализа данных типа времени жизни наиболее часто возникает необходимость обработки выборок, цензурированных справа. Появление цензурированных наблюдений является естественным и порождается спецификой проведения экспериментов.

Цензурированная справа выборка может быть представлена в виде

$$(X_1, \delta_1), (X_2, \delta_2), \dots, (X_n, \delta_n),$$

* Получена 10 августа 2012 г., после исправления – 20 сентября 2012 г.
Исследования выполнены при финансовой поддержке Министерства образования и науки Российской Федерации в рамках госзадания (проект 8.1274.2011) и ФЦП «Научные и научно-педагогические кадры инновационной России».

где $X_i = \min(T_i, C_i)$ – значение наблюдения, T_i – момент наступления системного события (отказа), C_i – момент цензурирования,

$$\delta_i = \begin{cases} 1, & \text{если } T_i < C_i, X_i = T_i \\ 0, & \text{если } T_i \geq C_i, X_i = C_i \end{cases}$$

– индикатор цензурирования, который равен единице, если i -е наблюдение полное (наблюдался отказ), нулю – если цензурированное. Выборка называется *случайно цензурированной* или *цензурированной III типа*, если T_i и C_i представляют собой независимые случайные величины, причем T_i принадлежит закону распределения вероятностей с функцией $F(x)$, а C_i – закону $F_C(x)$.

В настоящей работе исследуются модифицированные критерии согласия типа Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга, применяемые для проверки простых и сложных гипотез в условиях цензурирования. Модификации рассматриваемых критериев для проверки гипотез по выборкам с цензурированием I и II типа предложены в работах [1], [2], [3]. Результаты исследования распределений статистик таких модификаций критериев при проверке простых и сложных гипотез с использованием методики компьютерного моделирования представлены в [4]. Там же обсуждается влияние на распределения статистик критериев степени цензурирования и приводится сравнительный анализ мощности критериев относительно близких конкурирующих гипотез.

Модификации критериев Колмогорова и Крамера–Мизеса–Смирнова для проверки гипотез по случайно цензурированным выборкам предложены в работах [5], [6], [7], [8]. В данных модификациях при вычислении статистики критерия вместо эмпирической функции распределения используется непараметрическая оценка Каплана–Мейера, строящаяся по случайно цензурированной выборке. Верхние процентные точки предельных распределений статистик таких модифицированных критериев для проверки простых гипотез о согласии с законами нормальным, экспоненциальным и Вейбулла при заданном распределении моментов цензурирования приведены в [7].

Несмотря на то что модификациям критериев согласия для случайно цензурированных выборок посвящено достаточно много публикаций, применение данных критериев на практике вызывает серьезные затруднения. Во-первых, предельные распределения статистик рассматриваемых критериев зависят от распределения моментов цензурирования и неизвестны. Поэтому как при проверке простой, так и сложной гипотезы по случайно цензурированной выборке распределение статистики применяемого критерия при справедливости проверяемой гипотезы может быть найдено только в результате компьютерного моделирования. Во-вторых, в реальных ситуациях распределение моментов цензурирования, как правило, оказывается неизвестным. И возникает вопрос, каким образом моделировать распределение статистики критерия при неизвестном распределении моментов цензурирования?

В данной работе распределения статистик модифицированных критериев при проверке простых и сложных гипотез по случайно цензурированным данным при различных объемах выборок, степенях цензурирования и распределениях моментов цензурирования исследуются с использованием методов компьютерного моделирования. Предложен алгоритм моделирования распределений статистик рассматриваемых критериев в случае неизвестного распределения моментов цензурирования. Приводится пример проверки сложной гипотезы о согласии с использованием исследуемых критериев.

1. КРИТЕРИИ СОГЛАСИЯ ТИПА КОЛМОГорова, КРАМЕРА–МИЗЕСА–СМИРНОВА И АНДЕРСОНА–ДАРЛИНГА

При случайном цензурировании вместо эмпирической функции распределения $F_n(x)$ в критериях согласия Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга в [5], [6], [7], [8] предлагается использовать оценку Каплана–Мейера. Без потери общности будем

предполагать, что областью определения случайной величины является интервал $[0, \infty)$. Однако значения статистик рассматриваемых модифицированных критериев вычисляются на наблюдаемом интервале $[0, \tau]$, где τ – время последнего полного наблюдения. Обозначим через $a_1 < a_2 < \dots < a_k = \tau$, $k \leq n$ моменты времени, в которые были зафиксированы системные события $(X_i, \delta_i = 1)$. Тогда оценку Каплана–Мейера можно вычислить по формуле

$$\hat{F}_n(x) = \begin{cases} 0, & x < a_1, \\ 1 - \prod_{i: a_i \leq x} \left(1 - \frac{d_i}{r_i}\right), & x \geq a_1, \end{cases} \quad (1)$$

где $d_i = \sum_{j: X_j = a_i} \delta_j$, r_i – количество наблюдений, для которых $X_j \geq a_i$, $j = 1, \dots, n$.

В модифицированном критерии Колмогорова для случайно цензурированных выборок в качестве расстояния между эмпирическим и теоретическим законами распределения используется величина

$$D_n = \sup_{0 \leq x \leq \tau} |\hat{F}_n(x) - F(x; \theta)|, \quad (2)$$

где $\hat{F}_n(x)$ – оценка Каплана–Мейера, $F(x; \theta)$ – теоретическая функция распределения, соответствующая проверяемой гипотезе.

В модифицированном критерии Колмогорова будем использовать статистику вида

$$S_K^C = \frac{6nD_n + 1}{6\sqrt{n}} \quad (3)$$

с поправкой Большева, где $D_n = \max\{D_n^+, D_n^-\}$, $D_n^+ = \max_{1 \leq i \leq k} \{\hat{F}_n(a_i) - F(a_i; \theta)\}$,

$$D_n^- = \max_{1 \leq i \leq k} \{F(a_i; \theta) - \hat{F}_n(a_{i-1})\}.$$

В модифицированном критерии Крамера–Мизеса–Смирнова в качестве расстояния между распределениями используется величина

$$\omega^2 = \int_0^\tau (\hat{F}_n(x) - F(x; \theta))^2 dF(x; \theta).$$

Статистика модифицированного критерия Крамера–Мизеса–Смирнова с оценкой Каплана–Мейера имеет вид

$$S_\omega^C = \frac{n}{3} \cdot F(a_1; \theta) + n \cdot \sum_{j=1}^{k-1} \left[\hat{F}_n^2(a_j) (F(a_{j+1}; \theta) - F(a_j; \theta)) - \hat{F}_n(a_j) (F^2(a_{j+1}; \theta) - F^2(a_j; \theta)) + \frac{1}{3} (F^3(a_{j+1}; \theta) - F^3(a_j; \theta)) \right]. \quad (4)$$

В модифицированном критерии Андерсона–Дарлингга в качестве меры рассматривается величина

$$\Omega^2 = \int_0^\tau (\hat{F}_n(x) - F(x; \theta))^2 \frac{dF(x; \theta)}{F(x; \theta)(1 - F(x; \theta))}.$$

Соответственно, статистика модифицированного критерия Андерсона–Дарлинга принимает вид

$$S_{\Omega}^C = n \cdot \left\{ -F(a_1; \theta) + \sum_{j=1}^{k-1} \left[F(a_j; \theta) - F(a_{j+1}; \theta) + \hat{F}_n^2(a_j) (\ln F(a_{j+1}; \theta) - \ln F(a_j; \theta)) - \right. \right. \\ \left. \left. - (1 - \hat{F}_n(a_j))^2 (\ln(1 - F(a_{j+1}; \theta)) - \ln(1 - F(a_j; \theta))) \right] - \ln(1 - F(a_1; \theta)) \right\}. \quad (5)$$

Проверяемая гипотеза о согласии отвергается при больших значениях статистик. Аналитические выражения для распределений статистик рассматриваемых критериев неизвестны. Поэтому вычисление критических значений статистик (или достигнутых уровней значимости) при проверке гипотез с использованием данных критериев может опираться только на распределения статистик, полученные в результате статистического моделирования.

2. ИССЛЕДОВАНИЕ ОЦЕНОК КАПЛАНА–МЕЙЕРА

Распределения статистик рассматриваемых критериев согласия зависят от свойств оценок Каплана–Мейера, на основе которых вычисляются значения статистик. В первую очередь выясним, какое влияние оказывают на оценки Каплана–Мейера степень цензурирования (процент цензурированных наблюдений в выборке) и распределение моментов цензурирования.

В качестве примера распределения отказов $F(x)$ рассмотрим закон Вейбулла с функцией распределения

$$F(x; \theta) = W(\theta_1, \theta_2, \theta_3) = 1 - \exp \left(- \left(\frac{x - \theta_1}{\theta_2} \right)^{\theta_3} \right)$$

и значениями параметров $\theta_1 = 0, \theta_2 = 2, \theta_3 = 2$. В качестве распределений моментов цензурирования $F_C(x)$ выбраны два семейства:

- 1) семейство бета-распределений 1-го рода с функцией распределения

$$F(x; \theta) = B_I(\theta_1, \theta_2, \theta_3, \theta_4) = \frac{1}{B(\theta_3, \theta_4)} \int_0^{(x-\theta_1)/\theta_2} t^{\theta_3-1} (1-t)^{\theta_4-1} dt,$$

где $B(a, b)$ – бета-функция;

- 2) семейство распределений Вейбулла.

Значения параметров распределений моментов цензурирования были подобраны методами имитационного моделирования таким образом, чтобы средняя степень цензурирования была равна заданному значению. Полученные законы распределения моментов цензурирования приведены в табл. 1.

Таблица 1

Распределения моментов цензурирования

Средняя степень цензурирования, %	Распределение моментов цензурирования	
10	$B_1(0,7,1,81,1)$	$W(0,3,44,6,88)$
20	$B_1(0,7,1,19,1)$	$W(0,2,87,5,74)$
30	$B_1(0,7,1,1,24)$	$W(0,2,48,4,96)$
40	$B_1(0,7,1,1,83)$	$W(0,2,16,4,32)$
50	$B_1(0,7,1,2,58)$	$W(0,1,87,3,74)$
60	$B_1(0,7,1,3,58)$	$W(0,1,59,3,18)$
70	$B_1(0,7,1,5,01)$	$W(0,1,31,2,62)$
80	$B_1(0,7,1,7,36)$	$W(0,1,00,2,00)$

Теоретические функции распределения моментов цензурирования C_i , соответствующие закону Вейбулла и бета-распределению 1-го рода со значениями параметров, приведенными в табл. 1, представлены на рис. 1 и 2. На этих же рисунках отображена функция распределения Вейбулла, рассматриваемая в качестве функции распределения отказов T_i , $i = 1, 2, \dots, n$.

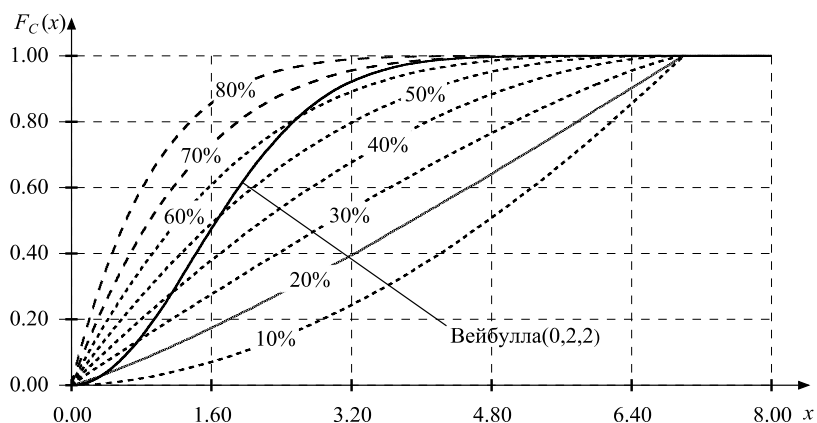


Рис. 1. Функция распределения отказов и функции распределения моментов цензурирования по законам Бета 1-го рода

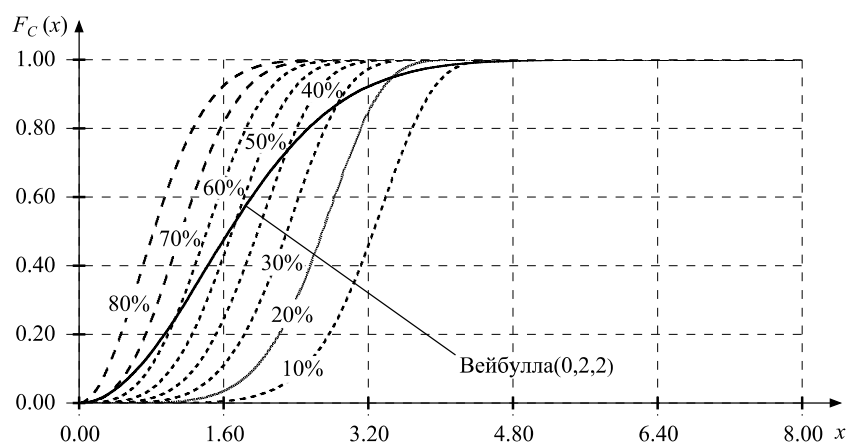


Рис. 2. Функция распределения отказов и функции распределения моментов цензурирования по законам Вейбулла

Как видим, взаимное расположение закона распределения отказов и распределений моментов цензурирования для разных семейств распределений оказывается различным. Следовательно, при одной и той же степени цензурирования расположение моментов цензурирования в вариационных рядах соответствующих выборок будет существенно отличаться. В частности, в случае принадлежности моментов цензурирования распределениям Вейбулла цензурированные наблюдения, как правило, оказываются в конце вариационного ряда моделируемых выборок. В случае же бета-распределений 1-го рода со значениями параметров, представленными в табл. 1, цензурированные наблюдения с большей вероятностью оказываются в начале вариационного ряда. Естественно, это будет отражаться на значениях статистик (3), (4), (5).

В табл. 2 представлены значения расстояния D_n между оценками Каплана–Мейера и теоретической функцией распределения Вейбулла, соответствующей истинному распределению отказов (при справедливости проверяемой гипотезы), усредненные по $N = 100\,000$ экспериментам.

Таблица 2

Отклонения оценок Каплана–Мейера от теоретической функции распределения

Объем выборки, n		100		200	
Семейство распределений $F_C(x)$		Вейбулла	Бета 1-го рода	Вейбулла	Бета 1-го рода
Средняя степень цензурирования, %	0	0.0854		0.0606	
	10	0.0865	0.0882	0.0616	0.0627
	20	0.0908	0.0935	0.0655	0.0664
	30	0.0975	0.0994	0.0724	0.0707
	40	0.1062	0.1077	0.0814	0.0768
	50	0.1164	0.1198	0.0922	0.0857
	60	0.1272	0.1369	0.1040	0.1000
	70	0.1383	0.1613	0.1176	0.1216
	80	0.1521	0.1899	0.1351	0.1539

Объем выборки, n	500		1000		2000	
Семейство распределений $F_C(x)$	Вейбулла	Бета 1-го рода	Вейбулла	Бета 1-го рода	Вейбулла	Бета 1-го рода
0	0.0385		0.0273		0.0193	
10	0.0392	0.0397	0.0279	0.0282	0.0197	0.0199
20	0.0430	0.0422	0.0311	0.0299	0.0231	0.0213
30	0.0496	0.0449	0.0374	0.0319	0.0290	0.0227
40	0.0584	0.0491	0.0460	0.0346	0.0369	0.0246
50	0.0691	0.0549	0.0557	0.0389	0.0463	0.0275
60	0.0813	0.0649	0.0677	0.0461	0.0571	0.0329
70	0.0954	0.0817	0.0806	0.0596	0.0691	0.0433
80	0.1143	0.1118	0.1006	0.0865	0.0871	0.0657

Как и следовало ожидать, отклонение оценок Каплана–Мейера от теоретической функции распределения отказов уменьшается с ростом объема выборки. В данном случае интересно различие в том, как уменьшается отклонение при разных распределениях моментов цензурирования. Например, когда в качестве $F_C(x)$ рассматривалось бета-распределение 1-го рода, среднее отклонение D_n уменьшается быстрее, чем в случае, когда моменты цензурирования принадлежат распределению Вейбулла и концентрируются в правой части области определения моментов отказа.

При отсутствии в выборке цензурированных наблюдений оценка Каплана–Мейера сводится к эмпирической функции распределения. В этом случае статистика $\sqrt{n}D_n$ при $n \rightarrow \infty$ подчиняется закону Колмогорова [9] и, начиная с некоторого n , математическое ожидание данной статистики становится практически независимым от объема выборки. На рис. 3 показаны изменения оценки математического ожидания статистики $\sqrt{n}D_n$ с ростом объема выборки при степени цензурирования около 50 % и различных законах распределения моментов цензурирования.

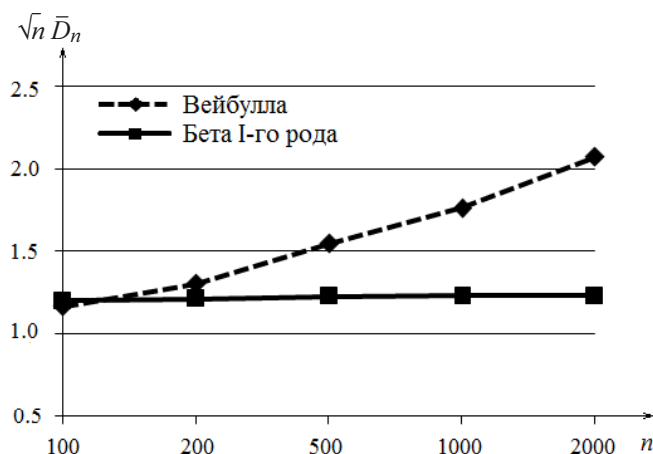


Рис. 3. Зависимость средних отклонений $\sqrt{n} \bar{D}_n$ от объема выборки при различных распределениях моментов цензурирования

Как видим на рис. 3, в случае принадлежности моментов цензурирования бетараспределению 1-го рода оценка математического ожидания практически не зависит от объема выборки. В случае же принадлежности моментов цензурирования распределению Вейбулла исследуемая величина с ростом объема выборки заметно увеличивается. Т. е., распределение моментов цензурирования существенно влияет на степень близости оценок Каплана–Мейера к истинной функции распределения отказов.

СПИСОК ЛИТЕРАТУРЫ

- [1] Barr D.M., Davidson T. A Kolmogorov-Smirnov test for censored samples. – Technometrics, 1973. – V.15. – № 4.
- [2] Pettitt A.N., Stephens M.A. Modified Cramer von Mises statistics for censored data // Biometrika. – 1976. – V. 63. – N. 2.
- [3] Мания Г.М. Статистическое оценивание распределений. – Тбилиси: Изд-во ТГУ, 1974. – 237 с.
- [4] Лемешко Б.Ю., Чимитова Е.В., Плешкова Т.А. Проверка простых и сложных гипотез о согласии по цензурированным выборкам // Научный вестник НГТУ. – 2010. – № 4(41). – С. 13–28.
- [5] Hjort N.L. On Inference in Parametric Survival Data // International Statistical Review. – 1992. – V.60. – № 3. – P. 355–387.
- [6] Koziol J.A., Green S.B. A Cramer-von Mises statistic for randomly censored data. Biometrika, 1976. – V. 63. – № 3. – P. 465–474.
- [7] Nair V. Plots and tests for goodness of fit with randomly censored data // Biometrika. – 1981. – V. 68. – P. 99–103.
- [8] Reineke D., Crown J. Estimation of Hazard, Density and Survival Functions for Randomly Censored Data // Journal of Applied Statistics. – 2004. – V.31. – № 10. – P. 1211–1225.
- [9] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. – 416 с.
- [10] Lee E.T., Wang J.W. Statistical methods for survival data analysis. – N.-J.: John Wiley & Sons, Inc., 2003. – 535 P.
- [11] Рыбалко В.В. Математические модели контроля надежности объектов энергетики. – СПб.: ГОУВПО СПбГТУРП., 2010. – 151 с.

Лемешко Борис Юрьевич, доктор технических наук, профессор кафедры прикладной математики Новосибирского государственного технического университета. Основное направление научных исследований – компьютерные технологии анализа данных и исследования статистических закономерностей. Имеет более 300 публикаций, в том числе 5 монографий. E-mail: lemeshko@fpm.ami.nstu.ru.

Чимитова Екатерина Владимировна, кандидат технических наук, доцент кафедры прикладной математики Новосибирского государственного технического университета. Основное направление научных исследований – статистические анализ данных типа времени жизни. Имеет более 50 публикаций. E-mail: ekaterina.chimitova@gmail.com.

Ведерникова Мария Александровна, аспирант кафедры прикладной математики Новосибирского государственного технического университета. Основное направление научных исследований – статистические методы анализа цензурированных данных. Имеет 10 публикаций. E-mail: vedernikova.m.a@gmail.com.

B.Yu. Lemeshko, E.V. Chimitova, M.A. Vedernikova

Modified goodness-of-fit tests of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling for randomly censored samples

The distributions of modified Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling goodness-of-fit test statistics have been studied by means of computer simulation methods for various sample sizes and distributions of censoring times. Testing simple and composite hypotheses has been considered. The algorithm for simulation of a randomly censored sample when the distribution of censoring times is unknown has been developed. The example of testing the composite goodness-of-fit hypothesis with considered modified tests for randomly censored data is given.

Key words: random censoring, modified goodness-of-fit tests of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling, Kaplan-Meier estimate.