*Научный вестник НГТУ. – 2013. – № 1(50)*

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.2

# Модифицированные критерии согласия Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга для случайно цензурированных выборок. Ч. 2*

**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 $\theta = (\theta_1, \theta_2, ...\theta_m)^T$ 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

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

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

* *Статья получена 10 августа 2012 г.*

*Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 14.B37.21.0860.*

*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 серьезнейшие 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 $\tau$ – время последнего полного наблюдения. Обозначим

через $a_1 < a_2 < ... < a_k = \tau$, $k \le n$, моменты времени, в которые были зафиксированы системные события $(X_i, \delta_i = 1)$. Тогда оценку Каплана–Мейера можно вычислить по формуле

$$\hat{F}_n(x) = \begin{cases} 0, \ x < a_1, \\ 1 - \prod\limits_{i: a_i \le x} \left(1 - \dfrac{d_i}{r_i}\right), \ x \ge a_1, \end{cases} \tag{1}$$

где $d_i = \sum\limits_{j: X_j = a_i} \delta_j$, $r_i$ – количество наблюдений, для которых $X_j \ge a_i$, $j = 1,...,n$.

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

$$D_n = \sup_{0 \le x \le \tau} \left| \hat{F}_n(x) - F(x;\theta) \right|, \tag{2}$$

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

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

с поправкой Большева, где $D_n = \max\{D_n^+, D_n^-\}$, $D_n^+ = \max\limits_{1 \le i \le k} \left\{ \hat{F}_n(a_i) - F(a_i;\theta) \right\}$, $D_n^- = \max\limits_{1 \le i \le k} \left\{ F(a_i;\theta) - \hat{F}_n(a_{i-1}) \right\}$.

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

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

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

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

$$- \hat{F}_n(a_j) \left( F^2(a_{j+1};\theta) - F^2(a_j;\theta) \right) + \frac{1}{3} \left( F^3(a_{j+1};\theta) - F^3(a_j;\theta) \right) \Big]. \tag{4}$$

В модифицированном критерии Андерсона–Дарлинга в качестве меры рассматривается величина

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

Соответственно, статистика модифицированного критерия Андерсона–Дарлинга принимает вид

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

$$\left.\left. - \left(1-\hat{F}_n\left(a_j\right)\right)^2 \left(\ln\left(1-F\left(a_{j+1};\theta\right)\right) - \ln\left(1-F\left(a_j;\theta\right)\right)\right) \right] - \ln\left(1-F\left(a_1;\theta\right)\right) \right\}. \qquad (5)$$

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 $F\left(x\right)$ рассмотрим закон Вейбулла с функцией распределения

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

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

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

$$F\left(x;\theta\right) = \beta_1\left(\theta_1,\theta_2,\theta_3,\theta_4\right) = \frac{1}{\beta\left(\theta_3,\theta_4\right)} \int_0^{(x-\theta_1)/\theta_2} t^{\theta_3-1}\left(1-t\right)^{\theta_4-1}\,dt,$$

где $\beta\left(a,b\right)$ – 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.

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 $T_i$, $i = 1,2,...,n$.

*Таблица 1*

**Распределения моментов цензурирования**

| Средняя степень цензурирования, % | Распределение моментов цензурирования | |
|---|---|---|
| 10 | $\beta_1(0,7,1.81,1)$ | $W(0,3.44,6.88)$ |
| 20 | $\beta_1(0,7,1.19,1)$ | $W(0,2.87,5.74)$ |
| 30 | $\beta_1(0,7,1,1.24)$ | $W(0,2.48,4.96)$ |
| 40 | $\beta_1(0,7,1,1.83)$ | $W(0,2.16,4.32)$ |
| 50 | $\beta_1(0,7,1,2.58)$ | $W(0,1.87,3.74)$ |
| 60 | $\beta_1(0,7,1,3.58)$ | $W(0,1.59,3.18)$ |
| 70 | $\beta_1(0,7,1,5.01)$ | $W(0,1.31,2.62)$ |
| 80 | $\beta_1(0,7,1,7.36)$ | $W(0,1.00,2.00)$ |

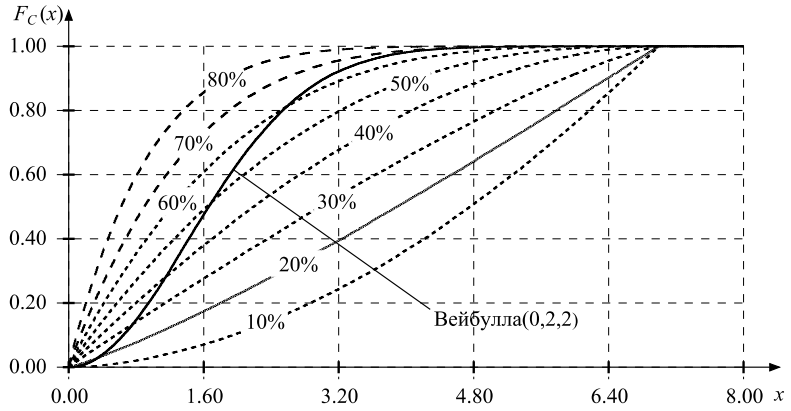


*Рис. 1.* Функция распределения отказов и функции распределения моментов цензурирования по законам бета 1-го рода

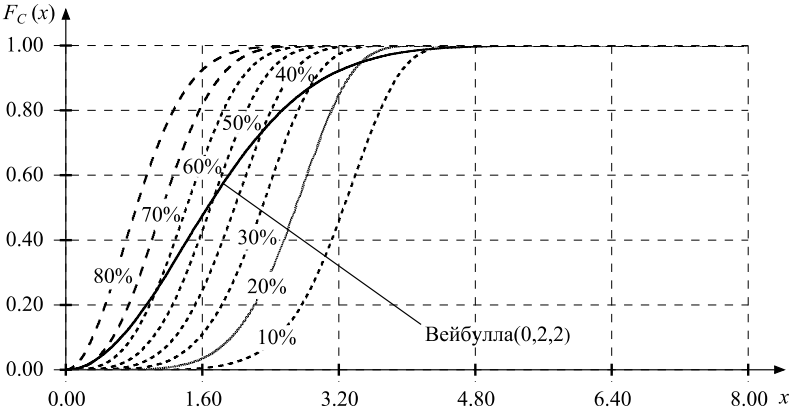


*Рис. 2.* Функция распределения отказов и функции распределения моментов цензурирования по законам Вейбулла

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 цензурирован-

ные 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 $N = 100000$ 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 $D_n$ уменьшалось быстрее, чем в случае, когда моменты цензурирования принадлежали распределению Вейбулла и концентрировались 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 $n \to \infty$ подчиняется закону Колмогорова [9] и, начиная с некоторого $n$, математическое ожидание данной статистики становится практически независящим 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.

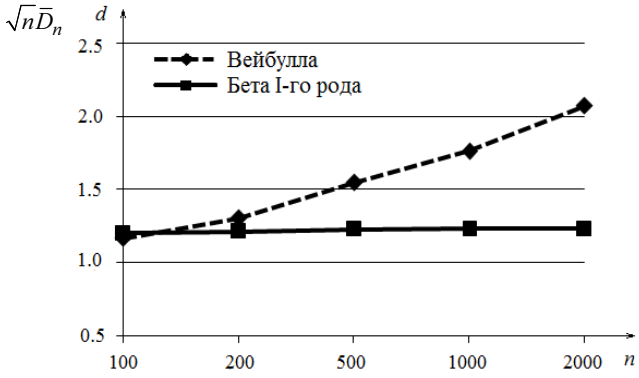


*Рис. 3.* Зависимость средних отклонений $\sqrt{n}\bar{D}_n$ от объема выборки при различных распределениях моментов цензурирования

Как видим на рис. 3, в случае принадлежности моментов цензурирования бета-распределению 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. То есть распределение моментов цензурирования существенно влияет на степень близости оценок Каплана–Мейера к истинной функции распределения отказов.

## 3. ИССЛЕДОВАНИЕ РАСПРЕДЕЛЕНИЙ СТАТИСТИК МОДИФИЦИРОВАННЫХ КРИТЕРИЕВ СОГЛАСИЯ

### 3.1. Исследование распределений статистик с ростом объема выборки

Методами компьютерного моделирования была исследована зависимость распределений статистик модифицированных критериев от объема выборок.

На рис. 4 представлены эмпирические распределения $G\left(S\middle|H_0\right)$ статистики модифицированного критерия Андерсона–Дарлинга при справедливости простой проверяемой гипотезы $H_0$ о принадлежности выборки распределению Вейбулла с параметрами (0, 2, 2). Распределения статистики получены при моделировании $N = 100000$ случайно цензурированных выборок в случае принадлежности моментов цензурирования бета-распределению 1-го рода с параметрами (0, 7, 1, 1.24) при средней степени цензурирования 30 %. В данном случае распределения статистики критерия практически не зависят от объема выборок $n$.

На рис. 5 показаны эмпирические распределения статистики критерия Андерсона–Дарлинга, полученные при тех же условиях проведения эксперимента, но в случае принадлежности моментов цензурирования распределению Вейбулла с параметрами (0, 2.48, 4.96). Средняя степень цензурирования также равна 30 %. Однако в данном случае мы видим существенную зависимость распределения статистики от $n$.

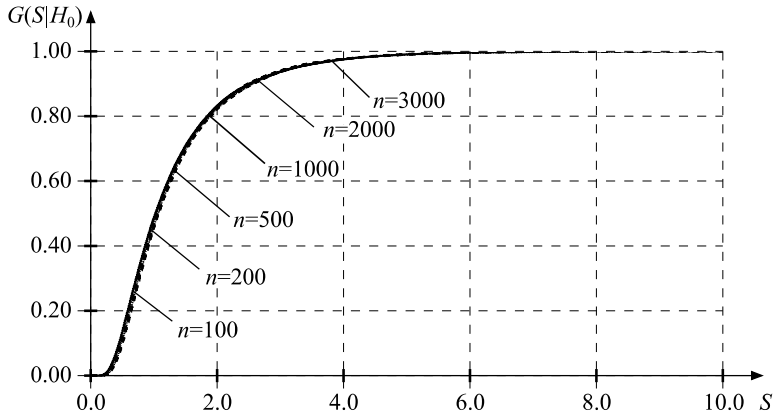


*Рис. 4*. Распределения статистики модифицированного критерия Андерсона–Дарлинга при проверке простой гипотезы в случае принадлежности моментов цензурирования бета-распределению 1-го рода при различных объемах выборок

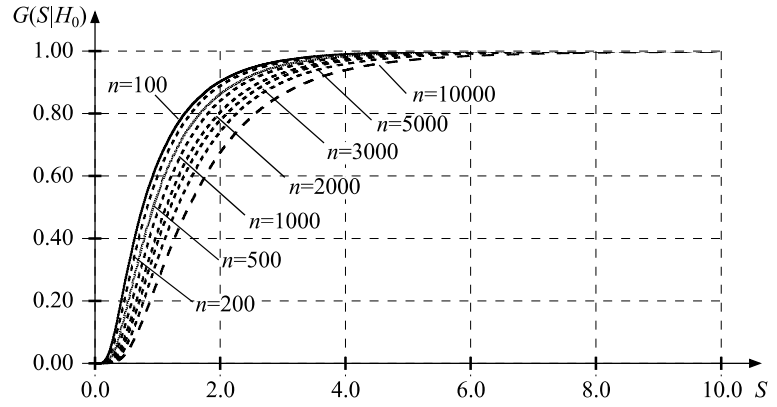


*Рис. 5*. Распределения статистики модифицированного критерия Андерсона–Дарлинга при проверке простой гипотезы в случае принадлежности моментов цензурирования распределению Вейбулла при различных объемах выборок

Таким образом, в общем случае распределения статистики модифицированного критерия Андерсона–Дарлинга зависят от объема выборки и от закона распределения моментов цензурирования. Если при степени цензурирования порядка 30 % и принадлежности моментов цензурирования бета-распределению 1-го рода распределения статистики с ростом объема выборок практически не меняются, то с увеличением степени цензурирования (более 60 %) распределения статистики становятся зависящими от объема выборок. И с ростом $n$ область определения статистики смещается в сторону больших значений (при любых из рассмотренных законах распределения моментов цензурирования). Такую зависимость распределений статистики можно легко объяснить, опираясь на результаты исследования математического ожидания величины $\sqrt{n}D_n$ (см. рис. 3).

Аналогичные результаты были получены для распределений статистик модифицированных критериев Колмогорова и Крамера–Мизеса–Смирнова при проверке как простых, так и сложных гипотез.

### 3.2. Исследование распределений статистик модифицированных критериев при различных степенях цензурирования

Результаты исследований распределений статистик модифицированных критериев от степени цензурирования демонстрируются на примере проверки простых гипотез о принадлежности выборок распределению Вейбулла при объеме выборок $n = 100$.

На рис. 6 представлены распределения $G\left(S\middle|H_0\right)$ статистики модифицированного критерия Андерсона–Дарлинга при проверке простой гипотезы в случае принадлежности моментов цензурирования бета-распределению 1-го рода для степеней цензурирования 10–70 % (см. табл. 1). Для сравнения на рисунке показано распределение $a2(S)$, являющееся предельным для критерия Андерсона–Дарлинга в случае полных выборок.

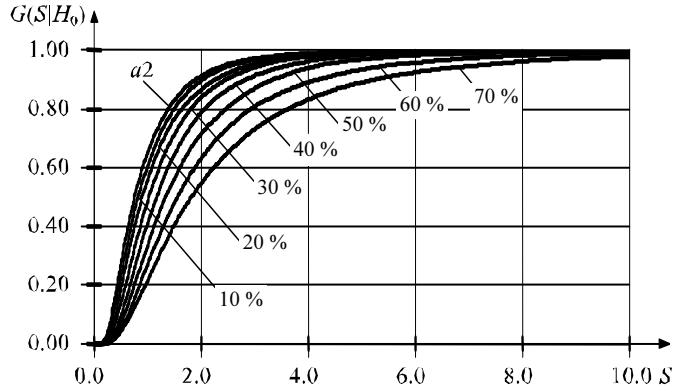


*Рис. 6*. Распределения статистики модифицированного критерия Андерсона–Дарлинга при проверке простой гипотезы в случае принадлежности моментов цензурирования бета-распределению 1-го рода при различных степенях цензурирования

Как можно видеть, с увеличением степени цензурирования распределения $G(S|H_0)$ статистики модифицированного критерия Андерсона–Дарлинга смещаются в область больших значений статистики. Понятно, что при проверке простых гипотез по случайно цензурированным выборкам распределение $a2(S)$ уже не является предельным распределением.

Вместе со степенью цензурирования на распределения статистик модифицированных критериев оказывает влияние и вид закона распределения моментов цензурирования $F_C\left(x\right)$. На рис. 7 показаны распределения статистики модифицированного критерия Андерсона–Дарлинга в случае принадлежности моментов цензурирования закону Вейбулла при степенях цензурирования 10–70 % (см. табл. 1). На рисунке для сравнения приведено также распределение $a2(S)$. Как видим, в данном случае зависимость распределения статистики от степени цензурирования выражена менее ярко.

При цензурировании I и II типа моделирование распределений статистик критериев согласия и построение для них приближенных моделей не вызывает принципиальных трудностей как при проверке простых, так и сложных гипотез [4]. Но при случайном цензурировании показанные зависимости распределений статистик модифицированных критериев согласия от объемов выборок и, главное, от закона распределения моментов цензурирования ставят под вопрос возможность построения приближенных моделей распределений статистик даже для проверки конкретной простой гипотезы. Проблема заключается в том, что в реальных приложениях закон распределения моментов цензурирования, как правило, неизвестен.

В общем случае при проверке сложных гипотез распределения статистик непараметрических критериев согласия зависят от закона $F_0\left(x;\theta\right)$, с которым проверяется согласие, от числа и типа оцениваемых параметров этого закона, от метода оценивания и, возможно, от значения

или значений конкретных параметров. При случайном цензурировании на это накладывается зависимость от закона распределения моментов цензурирования (и объема выборки).

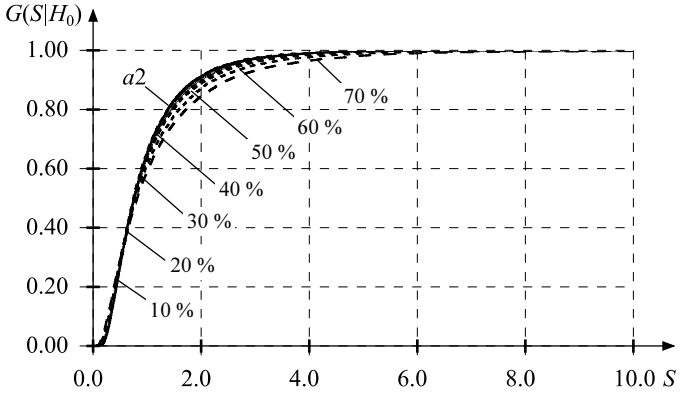


*Рис.* 7. Распределения статистики модифицированного критерия Андерсона–Дарлинга при проверке простой гипотезы в случае принадлежности моментов цензурирования распределению Вейбулла при различных степенях цензурирования

Таким образом, для проверки как простых, так и сложных гипотез с использованием модифицированных критериев согласия типа Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга необходимо иметь (знать) распределение статистики соответствующего критерия при справедливости проверяемой гипотезы (в конкретных условиях, соответствующих характеру регистрируемых наблюдений). Такие распределения могут быть найдены только в результате моделирования. Для моделирования распределений статистик необходимо подобрать распределение моментов цензурирования. Для построения параметрической модели распределения моментов цензурирования необходимо иметь некоторую априорную информацию, а после построения убедиться в адекватности этой модели. Поэтому для моделирования распределений статистик $G\left(S\middle|H_0\right)$ модифицированных критериев, необходимых при проверке гипотезы для определения достигнутого уровня значимости $P\left\{S \geq S^* \middle| H_0\right\} = 1 - G\left(S^* \middle| H_0\right)$, где $S^*$ – вычисленное по выборке значение статистики критерия, авторами был предложен и реализован следующий непараметрический алгоритм моделирования случайно цензурированных выборок.

### 3.3. Непараметрический алгоритм моделирования случайно цензурированной выборки

Для того чтобы смоделировать случайно цензурированную выборку в соответствии с механизмом цензурирования исходной (анализируемой, эталонной) выборки, необходимо выполнить следующую последовательность действий.

1. Смоделировать методом обратной функции полную выборку объемом $n$ по закону, соответствующему проверяемой гипотезе: $T_i = F^{-1}(\zeta_i;\theta)$, где $\zeta_i$ – псевдослучайная величина, равномерно распределенная на интервале (0,1), $i = 1,2,...,n$.

2. Инвертировать исходную цензурированную выборку, изменив значения индикаторов цензурирования $\delta_i$ на $1-\delta_i$.

3. Построить оценку Каплана–Мейера (1) функции распределения $\hat{F}_C\left(x\right)$ по инвертированной выборке.

4. Смоделировать $\xi_i$, равномерно распределенные на интервале (0,1), и вычислить значения $C_i$, $i = 1,2,...,n$:

a) если $\xi_i < \hat{F}_C(c_1)$, то $C_i = \dfrac{\xi_i \cdot c_1}{\hat{F}_C(c_1)}$;

b) если $\xi_i \in \left(\hat{F}_C(c_j), \hat{F}_C(c_{j+1})\right]$, то

$$C_i = c_j + \frac{\left(\xi_i - \hat{F}_C(c_j)\right)\cdot\left(c_{j+1} - c_j\right)}{\left(\hat{F}_C(c_{j+1}) - \hat{F}_C(c_j)\right)}, \quad j = 1,2,...,r;$$

c) если $\xi_i > \hat{F}_C(c_r)$, то $C_i = c_r + c_r\left(\xi_i - \hat{F}_C(c_r)\right)$, где $c_1,...,c_r$ – упорядоченные по возрастанию различные моменты цензурирования в исходной выборке, $r$ – количество различных моментов цензурирования в исходной выборке.

5. $X_i = \min(T_i, C_i)$, $\delta_i = 1\{T_i \le C_i\}$, $i = 1,2,...,n$.

Работоспособность алгоритма исследовалась проверкой однородности случайно цензурированных выборок, генерируемых параметрическим методом (с известными законами $F_0(x;\theta)$, $F_C(x)$) и в соответствии с предложенным алгоритмом (с известным законом $F_0(x;\theta)$ и механизмом цензурирования, извлекаемым из выборки, полученной параметрическим методом). Проверка гипотез об однородности получаемых выборок (при различной степени цензурирования) по критериям Гехана, Кокса–Мантела и логранговому [10] показала, что с высокими значениями достигнутых уровней значимости гипотеза об однородности не должна отклоняться. Достигнутые уровни значимости при проверке однородности генерируемых выборок объемом $n = 1000$ приведены в табл. 3.

*Таблица 3*

**Достигнутые уровни значимости при проверке однородности генерируемых выборок**

| Распределение моментов цензурирования | Критерий | Степень цензурирования, % | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 10 | 20 | 30 | 40 | 50 |
| Бета 1-го рода | Логранговый | 0.79 | 0.96 | 0.98 | 0.87 | 0.62 | 0.43 |
| | Гехана | 0.81 | 0.88 | 0.84 | 0.72 | 0.60 | 0.48 |
| | Кокса–Мантела | 0.87 | 0.91 | 0.90 | 0.82 | 0.62 | 0.58 |
| Вейбулла | Логранговый | 0.79 | 0.59 | 0.59 | 0.68 | 0.45 | 0.56 |
| | Гехана | 0.81 | 0.47 | 0.48 | 0.47 | 0.36 | 0.52 |
| | Кокса–Мантела | 0.87 | 0.51 | 0.47 | 0.53 | 0.49 | 0.59 |

Использование предложенного алгоритма моделирования случайно цензурированных выборок позволяет моделировать и исследовать распределения статистик модифицированных критериев согласия по цензурированным данным.

***Замечание.*** На распределения статистик существенное влияние оказывает степень цензурирования, которая, в свою очередь, определяется сочетанием распределений $F(x)$ и $F_C(x)$. Если распределение, соответствующее гипотезе $H_0$, достаточно близко к всегда неизвестному истинному распределению отказов, то средняя степень цензурирования в моделируемых выборках будет близка к степени цензурирования в исходной выборке, по которой проверяется согласие.

### 4. ПРИМЕР ПРОВЕРКИ СЛОЖНОЙ ГИПОТЕЗЫ О СОГЛАСИИ ПО СЛУЧАЙНО ЦЕНЗУРИРОВАННОЙ ВЫБОРКЕ

Рассмотрим пример проверки гипотезы о согласии с вероятностной моделью надежности по случайно цензурированной выборке, содержащей наработки до отказа одного из элементов газотурбогенераторов в течение продолжительного времени эксплуатации [11]. Объем выборки – 15 наблюдений, 6 из которых являются цензурированными, значения наблюдений – количество часов до наступления отказа. Данные представлены в табл. 4.

*Таблица 4*

**Вариационный ряд наработок элемента газотурбогенератора**

| Порядковый номер, $i$ | Значение наблюдения, $X_i$ | Индикатор цензурирования, $\delta_i$ |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 200 | 1 |
| 3 | 400 | 1 |
| 4 | 650 | 1 |
| 5 | 700 | 0 |
| 6 | 900 | 0 |
| 7 | 1200 | 0 |
| 8 | 1400 | 1 |
| 9 | 1550 | 1 |
| 10 | 1650 | 1 |
| 11 | 1800 | 0 |
| 12 | 1950 | 0 |
| 13 | 2000 | 1 |
| 14 | 3570 | 1 |
| 15 | 3700 | 1 |

В [11] в качестве модели надежности было предложено распределение Вейбулла. По исходной выборке получены оценки максимального правдоподобия: для параметра масштаба $\hat{\theta}_2 = 2286.4613$ и параметра формы $\hat{\theta}_3 = 1.5644$, параметр сдвига равен нулю.

На рис. 8 представлена оценка Каплана–Мейера, построенная по рассматриваемым данным, и теоретическая функция надежности, соответствующая распределению Вейбулла.

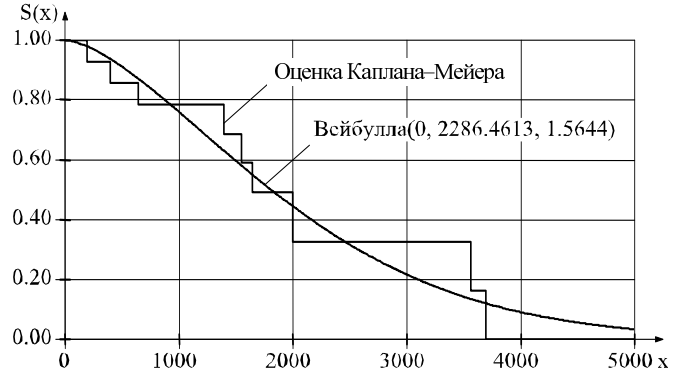


*Рис.* 8. Функция надежности Вейбулла и оценка Каплана–Мейера функции надежности, построенная по выборке

Проверим сложную гипотезу о согласии с распределением Вейбулла. Зафиксируем уровень значимости $\alpha = 0.1$.

Полученные по выборке отказов значения статистик модифицированных критериев Колмогорова (3), Крамера–Мизеса–Смирнова (4) и Андерсона–Дарлинга (5): $S_K^C = 0.7909$, $S_\omega^C = 0.0768$, $S_\Omega^C = 0.4818$.

Для вычисления достигнутых уровней значимости необходимо найти распределение статистик при справедливости сложной проверяемой гипотезы. Для этого в соответствии с предложенным алгоритмом моделировалось $N = 10^5$ случайно цензурированных выборок, по каждой выборке оценивались параметры распределения Вейбулла и вычислялись значения статистик рассматриваемых критериев. При этом не наблюдалось существенного отличия между количеством цензурированных наблюдений в генерируемых выборках (среднее число цензурированных наблюдений равно 7) и в исходной выборке (см. табл. 4).

На основе построенных эмпирических распределений статистик критериев Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга получены следующие достигнутые уровни значимости: $\alpha_K^C = 0.28$, $\alpha_\omega^C = 0.23$, $\alpha_\Omega^C = 0.24$. Поскольку вычисленные уровни значимости больше заданного $\alpha = 0.1$, то нет причин для отклонения проверяемой гипотезы.

## ЗАКЛЮЧЕНИЕ

Основное внимание в данной работе уделено проблеме применения модифицированных непараметрических критериев согласия в условиях, когда распределение моментов цензурирования неизвестно. Предложен алгоритм моделирования случайно цензурированных выборок, основанный на использовании оценки Каплана–Мейера для описания распределения моментов цензурирования.

Применение рассмотренных модифицированных критериев согласия для анализа случайно цензурированных выборок возможно при наличии программного обеспечения, позволяющего найти необходимое для проверки гипотезы распределение статистики критерия при справедливости проверяемой гипотезы в результате моделирования, осуществляемого в интерактивном режиме. При этом моделирование случайно цензурированных выборок с законом распределения моментов цензурирования, соответствующим анализируемой выборке при справедливости $H_0$, может осуществляться в соответствии с предложенным алгоритмом.

Предложенный алгоритм моделирования случайно цензурированных выборок реализован в программной системе статистического анализа данных типа времени жизни LiTiS (дистрибутив которой доступен по адресу http://amsa.conf.nstu.ru/amsa2011/Litis.msi). Данная программная система позволяет проверять простые и сложные гипотезы о согласии по цензурированным выборкам относительно широкого спектра законов распределения с использованием модифицированных критериев типа Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга.

Тем не менее, вопрос о возможности применения модифицированных критериев согласия в задачах идентификации вероятностных моделей надежности остается открытым. Для выбора закона распределения, наилучшим образом описывающего исходные данные, желательно использование критериев согласия, распределения статистик которых были бы в меньшей степени чувствительны к типу и степени цензурирования данных.

## СПИСОК ЛИТЕРАТУРЫ

[1] **Barr D.M.** A Kolmogorov-Smirnov test for censored samples / D.M. Barr, T. Davidson // Technometrics, 1973. – V. 15. № 4.

[2] **Pettitt A.N.** Modified Cramer von Mises statistics for censored data / A.N. Pettitt, M.A. Stephens // Biometrika. – 1976. – V. 63. №. 2.

[3] **Мания Г.М.** Статистическое оценивание распределений / Г.М. Мания. – Тбилиси: Изд-во ТГУ, 1974. – 237 с.

[4] **Лемешко Б.Ю.** Проверка простых и сложных гипотез о согласии по цензурированным выборкам / Б.Ю. Лемешко, Е.В. Чимитова, Т.А. Плешкова // Научный вестник НГТУ. – 2010. – № 4(41). – С.13–28.

[5] **Hjort N.L.** On Inference in Parametric Survival Data / N.L. Hjort // International Statistical Review. – 1992. – V. 60. № 3. – P. 355–387.

[6] **Koziol J.A.** A Cramer-von Mises statistic for randomly censored data / J.A. Koziol, S.B. Green // Biometrika. – 1976. – V. 63. № 3. – P. 465–474.

[7] **Nair V.** Plots and tests for goodness of fit with randomly censored data / V. Nair // Biometrika. – 1981. – V. 68. – P. 99–103.

[8] **Reineke D.** Estimation of Hazard, Density and Survival Functions for Randomly Censored Data / D. Reineke, J. Crown // Journal of Applied Statistics. – 2004. – V. 31. – № 10. – P. 1211–1225.

[9] **Большев Л.Н.** Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, 1983. – 416 с.

[10] **Lee E.T.** Statistical methods for survival data analysis / E.T. Lee, J.W. Wang. – NJ: John Wiley & Sons, Inc., 2003. – 535 P.

[11] **Рыбалко В.В.** Математические модели контроля надежности объектов энергетики / В.В. Рыбалко. – СПб.: ГОУВПО СПбГТУРП., 2010. – 151 с.

*Лемешко Борис Юрьевич*, доктор технических наук, профессор кафедры прикладной математики НГТУ. 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 НГТУ. 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. Основное направление научных исследований – статистические методы анализа цензурированных данных. E-mail: vedernikova.m.a@gmail.com

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

*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.