Научный вестник НГТУ. - 2010. - № 4(41). – С.13-28

УДК 519.2

Проверка простых и сложных гипотез о согласии по цензурированным выборкам*

Б.Ю. ЛЕМЕШКО, Е.В. ЧИМИТОВА, Т.А. ПЛЕШКОВА

В работе рассматриваются вопросы применения критериев согласия для проверки простых и сложных гипотез по односторонне цензурированным выборкам. Методами компьютерного моделирования исследуются распределения статистик критериев Реньи, Колмогорова, Крамера-Мизеса-Смирнова, Андерсона-Дарлинга при различных объемах выборок и степени цензурирования. Для случая проверки простых гипотез найдены минимальные объемы выборок, начиная с которых распределения статистики типа Колмогорова хорошо согласуются с предельным законом, построены таблицы верхних процентных точек для статистик критериев типа Колмогорова, Крамера-Мизеса-Смирнова, Андерсона-Дарлинга. Для ситуации проверки сложных гипотез относительно законов Вейбулла и логнормального с оцениванием неизвестных параметров методом максимального правдоподобия для перечисленных критериев построены таблицы верхних процентных точек распределений статистик.

Ключевые слова: цензурированная выборка, критерии согласия типа Реньи, Колмогорова, Крамера-Мизеса-Смирнова, Андерсона-Дарлинга

1. ВВЕДЕНИЕ

Практически в любой сфере научной деятельности, связанной с регистрацией наблюдений, возникает необходимость в статистической обработке полученных данных. Достоверность результатов статистического анализа в первую очередь зависит от степени адекватности выбранной модели анализируемым данным. Поэтому обязательным этапом анализа является проверка гипотезы о согласии результатов наблюдений случайной величины с выбранным теоретическим распределением.

Для проверки гипотезы о согласии в соответствии с выбранным критерием вычисляется значение статистики критерия как некоторой функции от выборки и закона распределения, с которым проверятся согласие. Для используемого критерия должно быть известно (предельное) распределение статистики при условии справедливости проверяемой (нулевой) гипотезы, либо известны верхние процентные точки распределения статистики. Тогда на основании вычисленного значения статистики можно принять решение отклонить или не отклонять проверяемую гипотезу.

При анализе величин типа “времени жизни” в задачах теории надежности или в медицинских и биологических исследованиях, как правило, сталкиваются с задачами обработки цензурированных выборок. Появление цензурированных выборок в этих задачах порождается спецификой проведения экспериментов и условиями регистрации наблюдений, а интерес к таким задачам постоянно растет. При проверке гипотез по цензурированным данным можно использовать критерии типа Реньи [1, 2,3], Колмогорова [3, 4], Андерсона-Дарлинга [5] или Крамера-Мизеса-Смирнова [5].

С применением непараметрических критериев согласия в случае сложных гипотез существуют проблемы даже при проверке гипотез по полным данным [6, 7, 8, 9], так как непараметрические критерии согласия теряют свойство “свободы от распределения”. При проверке сложных гипотез, когда оценки параметров закона вычисляются по тем же выборкам, распределения статистик непараметрических критериев согласия зависят от закона распределения, с которым проверяется согласие,

* *Статья получена 1 сентября 2010 г.*

от числа и типа оцениваемых параметров, в некоторых случаях, от конкретных значений параметров. В случае цензурированных данных проблемы возникают не только при сложных гипотезах, но и при простых, так как распределения статистик зависят от типа и степени цензурирования [10, 11].

Предельное распределение статистики Реньи, в отличие от распределений статистик Колмогорова, Крамера-Мизеса-Смирнова и Андерсона-Дарлинга, в случае проверки простых гипотез не зависит от степени цензурирования, и по идее, таким критерием удобнее пользоваться на практике. Однако вопрос о том, насколько хорошо распределения статистик Реньи согласуются с соответствующими предельными законами при ограниченных объемах выборок, до сих пор не исследовался. Неизвестно и то, насколько быстро сходятся к своим предельным законам распределения статистики критерия Колмогорова в случае цензурированных выборок.

Цель данной работы заключалась: в исследовании влияния степени цензурирования на распределения статистик непараметрических критериев согласия; в исследовании области корректного применения критериев при проверке простых и сложных гипотез; в построении таблиц верхних процентных точек для распределений статистик непараметрических критериев; в сравнительном анализе мощности непараметрических критериев согласия при близких конкурирующих гипотезах. Исследования проводились с использованием развиваемой методики компьютерного моделирования и анализа статистических закономерностей.

2. ОЦЕНКА ПАРАМЕТРОВ ПО ЦЕНЗУРИРОВАННЫМ ДАННЫМ

Введем основные обозначения.

Выборка называется *цензурированной справа и (или) слева типа I* в точке $x_{(r)}$ и (или) $x_{(l)}$ соответственно, если наблюдаются лишь те члены независимой выборки $X_1, X_2, ..., X_n$, значения которых лежат левее $x_{(r)}$ и (или) правее $x_{(l)}$.

Выборка называется *цензурированной справа и (или) слева типа II*, если наблюдаются, соответственно, лишь $n - n_r$ наименьших и (или) $n - n_l$ наибольших членов вариационного ряда, построенного по выборке $X_1, X_2, ..., X_n$, $n_r + n_l < n$.

В случае цензурирования I типа фиксируются вероятности попадания в интервалы цензурирования $(-\infty, x_{(l)})$ и $(x_{(r)}, +\infty)$, а n_l и n_r обозначают случайное число наблюдений, попавших в интервалы цензурирования. При II типе количество цензурированных наблюдений n_l и n_r известно, а граничные точки $x_{(l)}$ и $x_{(r)}$ случайны. В таком случае в качестве $x_{(l)}$ обычно выбирают наименьшее наблюдаемое значение, а в качестве $x_{(r)}$ – наибольшее.

Степенью цензурирования a будем называть вероятность попадания в интервал цензурирования в случае цензурирования I типа, или отношение количества цензурированных наблюдений к полному объему выборки в случае цензурирования II типа.

Как упоминалось выше, проблемы при проверке сложных гипотез возникают, при вычислении статистик критериев с использованием оценок параметров закона, полученным по тем же самым выборкам. При этом распределения статистик зависят от метода оценивания.

В случае цензурированных данных можно использовать различные методы оценивания. Но наиболее эффективным и универсальным по отношению к форме представления выборочных данных является метод максимального правдоподобия. Оценка максимального правдоподобия (ОМП) неизвестного параметра по цензурированной слева и справа выборке получается в качестве решения системы уравнений правдоподобия

$$n_r \frac{\partial \ln P_1(\theta)}{\partial \theta_l} + \sum_{j=1}^{n-(n_l+n_r)} \frac{\partial \ln f(x_j, \theta)}{\partial \theta_l} + n_l \frac{\partial \ln P_3(\theta)}{\partial \theta_l} = 0,\ l = \overline{1,m}, \quad (1)$$

где $P_1(\theta) = \int_{-\infty}^{x_{(r)}} f(x,\theta)dx$, $P_3(\theta) = \int_{x_{(l)}}^{\infty} f(x,\theta)dx$. В случае цензурирования только справа (только слева) в выражении исчезает первое (третье) слагаемое.

3. КРИТЕРИИ СОГЛАСИЯ ПО ЦЕНЗУРИРОВАННЫМ ДАННЫМ

Обозначим через $x_1, ..., x_n$ упорядоченные по возрастанию выборочные значения.

Критерий Реньи. Двусторонняя статистика критерия Реньи в случае цензурирования слева задается выражением [3]

$$S_R^c = \sqrt{\frac{na}{1-a}} \cdot \sup_{F(x) \ge a} \frac{|F_n(x) - F(x)|}{F(x)},$$

а в случае цензурирования справа –

$$S_R^c = \sqrt{\frac{na}{1-a}} \cdot \sup_{F(x) \le 1-a} \frac{|F_n(x) - F(x)|}{1 - F(x)},$$

где $a \in (0,1)$ – степень цензурирования. Для этой статистики при справедливости простой проверяемой гипотезы имеет место предельное соотношение:

$$\lim_{n\to\infty} P\ \{S_R^c < S\} = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{(2k+1)^2 \pi^2}{8S^2}\right\} = L(S). \quad (2)$$

Критерий Колмогорова. Статистику Колмогорова вычисляют следующим образом [1]

$$S_K^c = \frac{6nD_n + 1}{6\sqrt{n}},$$

где $D_n = \max\{D_n^+, D_n^-\}$ и $D_n^+ = \max_{n_l < i \le n} \left\{\frac{i}{n} - F(x_i)\right\}$, $D_n^- = \max_{n_l < i \le n} \left\{F(x_i) - \frac{i-1}{n}\right\}$ – в случае цензурирования слева, $D_n^+ = \max_{1 \le i \le n-n_r} \left\{\frac{i}{n} - F(x_i)\right\}$, $D_n^- = \max_{1 \le i \le n-n_r} \left\{F(x_i) - \frac{i-1}{n}\right\}$ – в случае цензурирования справа, где n_l и n_r – количество наблюдений, попавших в левый или правый интервал цензурирования, соответственно.

Предельное соотношение имеет вид

$$P\ \{S_K^c < S\} = \sum_{i=-\infty}^{+\infty} (-1)^i \exp(-2i^2 S^2) \cdot P\left\{\left|X - 2iS\sqrt{\frac{a}{1-a}}\right| < \frac{S}{\sqrt{a - a^2}}\right\} = K_a^c(S), \quad (3)$$

где X – случайная величина, подчиняющаяся стандартному нормальному закону. При $a = 0$ предельное распределение статистики S_K^c совпадает с классическим распределением Колмогорова (для случая полной выборки):

$$K(S) = \sum_{i=-\infty}^{+\infty} (-1)^i \exp(-2i^2 S^2) .$$

Критерий Крамера-Мизеса-Смирнова. Статистика критерия Крамера-Мизеса-Смирнова для I и II типов цензурированных выборок имеет вид:
при цензурировании справа –

$$S_\omega^c = \frac{1}{12n} + \sum_{i=1}^{n-n_r} \left[F\ x_i\ - \frac{2i-1}{2n} \right]^2 ;$$

при цензурировании слева –

$$S_\omega^c = \frac{1}{12n} + \sum_{i=n_l+1}^{n} \left[F\ x_i\ - \frac{2i-1}{2n} \right]^2 ,$$

где n_l и n_r - количество наблюдений, попавших в левый или правый интервал цензурирования соответственно.

Критерий Андерсона-Дарлинга. Статистика критерия Андерсона-Дарлинга для цензурированных выборок I и II типов имеет вид:
при цензурировании справа –

$$S_\Omega^c = \sum_{i=1}^{n-n_r-1} \left[F\ x_i\ - F\ x_{i+1}\ + \left(\frac{n-i}{n}\right)^2 \ln\left(\frac{1-F\ x_i}{1-F\ x_{i+1}}\right) + \left(\frac{i}{n}\right) \ln\left(\frac{F\ x_{i+1}}{F\ x_i}\right) \right] -$$
$$- \ln\ 1-F\ x_1\ - F\ x_1\ ;$$

при цензурировании слева –

$$S_\Omega^c = \sum_{i=n_l+1}^{n-1} \left[F\ x_i\ - F\ x_{i+1}\ + \left(\frac{n-i}{n}\right)^2 \ln\left(\frac{1-F\ x_i}{1-F\ x_{i+1}}\right) + \left(\frac{i}{n}\right)^2 \ln\left(\frac{F\ x_{i+1}}{F\ x_i}\right) \right] -$$
$$- 1 - \ln\ F\ x_n\ + F\ x_n\ .$$

Для критериев Крамера-Мизеса-Смирнова и Андерсона-Дарлинга не существует аналитических предельных соотношений для распределений статистик по цензурированным данным.

4. РАСПРЕДЕЛЕНИЯ СТАТИСТИК КРИТЕРИЕВ ПРИ ПРОВЕРКЕ ПРОСТЫХ ГИПОТЕЗ

Исследование распределений статистик проводилось с использованием развиваемой методики компьютерного моделирования и анализа статистических закономерностей. Методика позволяет быстро и не менее точно, чем с использованием строгого математического аппарата, прослеживать статистические закономерности.

Предельное распределение статистики критерия Реньи не зависит от степени цензурирования, однако отсутствует информация о том, насколько хорошо распределение статистики согласуется с предельным законом при ограниченных объемах выборок.

Результаты статистического моделирования показали, что распределения статистик критерия Реньи существенно зависят от степени и структуры цензурирования. Например, на рис. 1 приведены эмпирические распределения статистики критерия для случая проверки простой гипотезы относительно экспоненциального закона при степени цензурирования $a = 0.9$.

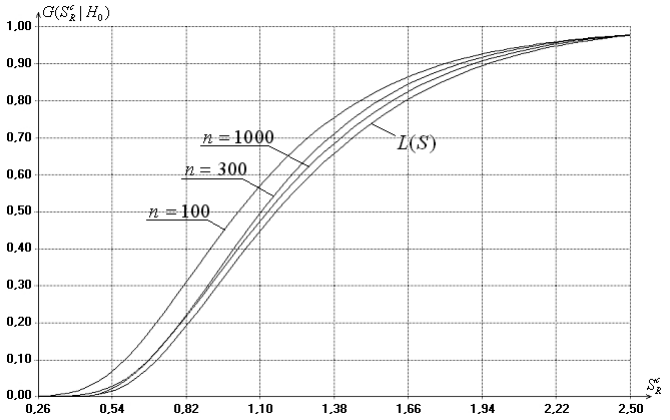

Рис. 1. Распределения статистики Реньи при цензурировании слева, $a = 0.9$

Можно отметить, что распределения статистики очень медленно сходятся к предельному и достаточно близкими к нему оказываются при полных значениях объема выборки $n > 1000$. При меньших n распределения статистики существенно отличаются от предельного.

На рис. 2 показана зависимость распределений статистики от степени цензурирования. На рисунке представлены распределения статистики при $n = 100$ и различной величине степени цензурирования. Как следует из картины, представленной на рисунке, наилучшее согласие с предельным распределением достигается при 50% наблюдаемой области (при $a = 0.5$), а при малой или, наоборот, высокой степени цензурирования распределения статистики существенно отличаются от предельного. То есть, применение критерия Реньи и использование предельного распределения его статистики при объемах выборок меньших 1000 и при степени цензурирования большей или меньшей 50% может приводить к некорректным выводам. Следовательно, *использование критерия Реньи в приложениях нецелесообразно*.

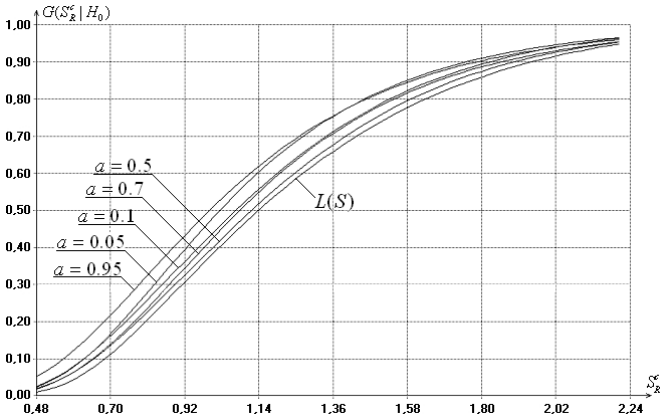

Рис. 2. Распределения статистики Реньи в зависимости от величины наблюдаемой области, $n = 100$, цензурирование слева I типа

Распределения статистики критерия Колмогорова исследовались при различных объемах выборок и различной степени цензурирования. Исследования показали, что уже при потенциальном объеме выборки $n = 30$ в случае степени цензурирования $a \le 0.5$ эмпирические распределения статистики Колмогорова хорошо согласуются с

соответствующими предельными. При больших степенях цензурирования требуются большие объемы выборок для достижения согласия эмпирических и предельных распределений Колмогорова. В качестве примера на рис. 3 приведены распределения статистики Колмогорова по цензурированным выборкам при проверке простой гипотезы о согласии с законом Вейбулла. На рисунке показаны полученные в результате моделирования распределения статистики $G(S_k^c | H_0)$ при справедливости проверяемой гипотезы H_0 при полном объеме выборок $n = 50$ и соответствующие предельные распределения статистики для степени цензурирования $a = 0.1,\ 0.5,\ 0.7,\ 0.9$.

В результате исследования распределений статистики Колмогорова найдены минимальные объемы выборок, при которых достигается хорошее согласие распределения статистики с соответствующим предельным законом для различных значений a степени цензурирования, которые представлены в таблице 1.

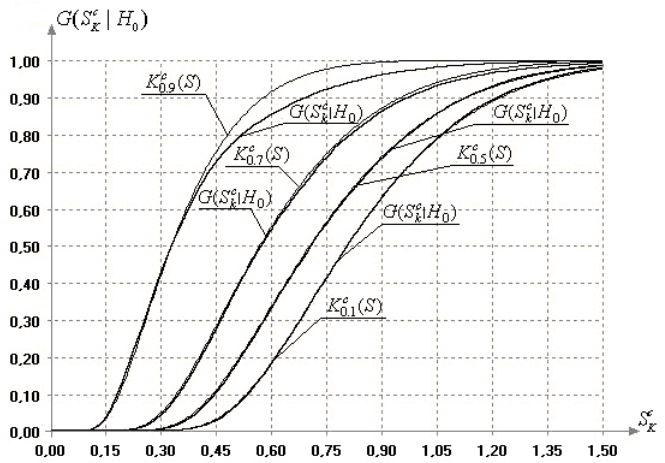

Рис. 3. Распределения статистики Колмогорова при различной степени цензурирования, $n = 50$, цензурирование слева II типа

Таблица 1. Требуемые минимальные объемы выборок для применения критерия Колмогорова

Степень цензурирования	Вид цензурирования			
	Слева, I тип	Слева, II тип	Справа, I тип	Справа, II тип
$a = 0.1$	20	20	20	20
$a = 0.2$	20	20	20	20
$a = 0.3$	20	20	20	30
$a = 0.4$	20	30	30	30
$a = 0.5$	30	30	30	30
$a = 0.6$	30	30	40	40
$a = 0.7$	40	40	40	50
$a = 0.8$	50	60	60	70
$a = 0.9$	100	250	110	140

Как было сказано выше, для критериев Крамера-Мизеса-Смирнова и Андерсона-Дарлинга отсутствуют аналитические предельные соотношения для распределений статистик по цензурированным данным. В случае проверки простых

гипотез распределения статистик непараметрических критериев согласия не зависят от законов распределений.

В данной работе методами компьютерного моделирования построены таблицы верхних процентных точек распределений статистик непараметрических критериев согласия для случая проверки простых гипотез. Процентные точки были получены по смоделированным выборкам статистик объемом $N = 10^6$. Значения статистик критериев вычислялись по выборкам псевдослучайных величин объемом $n = 10^3$.

Верхние процентные точки для предельных распределений статистики критерия Колмогорова при проверке простых гипотез представлены в таблице 2, для критерия Крамера-Мизеса-Смирнова – в таблице 3, для критерия Андерсона-Дарлинга – в таблице – 4. Гипотеза о согласии отвергается, если достигнутый уровень значимости меньше заданного или полученное по выборке значение статистики меньше критического значения, найденного по таблицам верхних процентных точек.

Таблица 2. Верхние процентные точки для статистики критерия Колмогорова

Степень цензурирования	Уровень значимости α		
	0.01	0.05	0.1
$a = 0$	1.628	1.358	1.224
$a = 0.1$	1.628	1.358	1.224
$a = 0.2$	1.621	1.347	1.209
$a = 0.3$	1.621	1.347	1.209
$a = 0.4$	1.600	1.321	1.181
$a = 0.5$	1.551	1.273	1.133
$a = 0.6$	1.467	1.198	1.062
$a = 0.7$	1.342	1.087	0.960
$a = 0.8$	1.151	0.927	0.815
$a = 0.9$	0.851	0.682	0.599

Таблица 3. Верхние процентные точки для статистики критерия Крамера-Мизеса-Смирнова

Степень цензурирования	Уровень значимости α		
	0.01	0.05	0.1
$a = 0$	0.741	0.463	0.348
$a = 0.1$	0.743	0.459	0.347
$a = 0.2$	0.705	0.433	0.323
$a = 0.3$	0.637	0.389	0.289
$a = 0.4$	0.545	0.33	0.242
$a = 0.5$	0.429	0.26	0.19
$a = 0.6$	0.311	0.186	0.136
$a = 0.7$	0.192	0.116	0.084
$a = 0.8$	0.097	0.057	0.041
$a = 0.9$	0.027	0.016	0.011

В случае проверки простой гипотезы значения процентных точек не зависят от вида цензурирования, однако зависят от степени цензурирования. При увеличении степени цензурирования a значения процентных точек уменьшаются. Построенные таблицы процентных точек могут использоваться при проверке простых гипотез.

Таблица 4. Верхние процентные точки для статистики критерия Андерсона-Дарлинга

Степень цензурирования	Уровень значимости α		
	0.01	0.05	0.1
$a = 0$	3.876	2.500	1.936
$a = 0.1$	3.737	2.385	1.831
$a = 0.2$	3.452	2.183	1.661
$a = 0.3$	3.118	1.946	1.477
$a = 0.4$	2.745	1.701	1.281
$a = 0.5$	2.348	1.428	1.071
$a = 0.6$	1.902	1.178	0.877
$a = 0.7$	1.454	0.892	0.666
$a = 0.8$	0.981	0.606	0.449
$a = 0.9$	0.507	0.302	0.225

5. ИССЛЕДОВАНИЕ РАСПРЕДЕЛЕНИЙ СТАТИСТИК КРИТЕРИЕВ СОГЛАСИЯ ПРИ ПРОВЕРКЕ СЛОЖНЫХ ГИПОТЕЗ

Чаще всего на практике приходится иметь дело с проверкой сложных гипотез, когда в качестве параметров исследуемого распределения берут оценки параметров, полученные по тем же самым данным. В случае проверки сложных гипотез распределения статистик зависят от вида закона, с которым проверяется согласие, от числа оцениваемых параметров, от вида параметров, метода оценивания, а в случае цензурированных наблюдений – еще и от степени и структуры цензурирования. Наиболее универсальным по отношению к форме представления данных является метод максимального правдоподобия. Оценки максимального правдоподобия по цензурированным данным являются асимптотически эффективными и несмещенными. Однако при ограниченных объемах выборок и больших степенях цензурирования, как показано в [10, 12, 13], ОМП оказываются смещенными, а их распределения – асимметричными. Это следует учитывать при проверке сложных гипотез.

В настоящей работе построены таблицы верхних процентных точек распределений статистик непараметрических критериев согласия при проверке сложных гипотез относительно законов распределения Вейбулла и логарифмически нормального (при оценке двух параметров закона методом максимального правдоподобия). Таблицы построены по смоделированным выборкам статистик объемом $N = 10^5$. При этом значения статистик критериев вычислялись по выборкам псевдослучайных величин с полным объемом $n = 10^3$. В полученных таблицах наибольшая рассматриваемая степень цензурирования 60%.

Верхние процентные точки для предельных распределений статистик непараметрических критериев при проверке сложных гипотез о согласии с законом Вейбулла представлены в таблицах 5-7, о согласии с логарифмически нормальным законом – в таблицах 8-10.

Из таблиц процентных точек видно, что характер изменения значений процентных точек зависит от вида цензурирования: при цензурировании слева значения возрастают, при цензурировании справа – убывают. Это связано с тем, что с увеличением степени цензурирования оценки параметров становятся все более смещенными и направление смещения оценок (влево или вправо) зависит от того, является ли выборка цензурированной слева или справа [10].

Таблица 5. Верхние процентные точки распределения статистики критерия Колмогорова при проверке сложных гипотез о согласии с законом Вейбулла

Степень цензурирования	Уровень значимости α					
	Цензурирование справа			Цензурирование слева		
	0.01	0.05	0.1	0.01	0.05	0.1
$a = 0$	1.039	0.895	0.825	1.039	0.895	0.825
$a = 0.1$	1.022	0.875	0.805	1.211	1.060	0.984
$a = 0.2$	0.993	0.849	0.78	1.473	1.328	1.258
$a = 0.3$	0.953	0.815	0.747	1.723	1.583	1.514
$a = 0.4$	0.904	0.771	0.706	1.913	1.775	1.699
$a = 0.5$	0.844	0.717	0.657	2.034	1.894	1.82
$a = 0.6$	0.771	0.654	0.597	2.106	1.962	1.876

Таблица 6. Верхние процентные точки распределения статистики критерия Крамера-Мизеса-Смирнова при проверке сложных гипотез о согласии с законом Вейбулла

Степень цензурирования	Уровень значимости α					
	Цензурирование справа			Цензурирование слева		
	0.01	0.05	0.1	0.01	0.05	0.1
$a = 0$	0.174	0.124	0.102	0.174	0.124	0.102
$a = 0.1$	0.159	0.111	0.091	0.258	0.194	0.166
$a = 0.2$	0.138	0.095	0.078	0.423	0.345	0.308
$a = 0.3$	0.114	0.078	0.064	0.589	0.496	0.456
$a = 0.4$	0.089	0.061	0.050	0.684	0.590	0.545
$a = 0.5$	0.067	0.045	0.036	0.701	0.611	0.565
$a = 0.6$	0.046	0.031	0.025	0.676	0.612	0.547

Таблица 7. Верхние процентные точки распределения статистики критерия Андерсона-Дарлинга при проверке сложных гипотез о согласии с законом Вейбулла

Степень цензурирования	Уровень значимости α					
	Цензурирование справа			Цензурирование слева		
	0.01	0.05	0.1	0.01	0.05	0.1
$a = 0$	1.039	0.756	0.638	1.039	0.756	0.638
$a = 0.1$	0.868	0.624	0.520	1.267	0.971	0.844
$a = 0.2$	0.745	0.526	0.438	1.917	1.572	1.412
$a = 0.3$	0.630	0.444	0.368	2.551	2.158	1.966
$a = 0.4$	0.522	0.371	0.305	2.973	2.543	2.334
$a = 0.5$	0.428	0.301	0.247	3.251	2.830	2.618
$a = 0.6$	0.337	0.235	0.193	3.710	3.288	3.076

Таблица 8. Верхние процентные точки распределения статистики критерия Колмогорова при проверке сложных гипотез о согласии с логарифмически нормальным законом

Степень цензурирования	Уровень значимости α					
	Цензурирование справа			Цензурирование слева		
	0.01	0.05	0.1	0.01	0.05	0.1
$a = 0$	1.601	1.331	1.192	1.601	1.331	1.192
$a = 0.1$	1.594	1.323	1.183	1.632	1.37	1.239
$a = 0.2$	1.591	1.318	1.177	1.601	1.331	1.192
$a = 0.3$	1.587	1.309	1.166	1.611	1.339	1.194
$a = 0.4$	1.571	1.286	1.141	1.810	1.539	1.403
$a = 0.5$	1.512	1.230	1.086	2.266	2.003	1.864
$a = 0.6$	1.393	1.122	0.990	2.925	2.670	2.534

Таблица 9. Верхние процентные точки распределения статистики критерия Крамера-Мизеса-Смирнова при проверке сложных гипотез о согласии с логарифмически нормальным законом

Степень цензурирования	Уровень значимости α					
	Цензурирование справа			Цензурирование слева		
	0.01	0.05	0.1	0.01	0.05	0.1
$a = 0$	0.682	0.417	0.305	0.682	0.417	0.305
$a = 0.1$	0.720	0.439	0.322	0.729	0.458	0.344
$a = 0.2$	0.685	0.416	0.303	0.682	0.417	0.305
$a = 0.3$	0.612	0.370	0.269	0.669	0.406	0.296
$a = 0.4$	0.509	0.303	0.219	0.847	0.550	0.421
$a = 0.5$	0.373	0.221	0.160	1.326	0.982	0.822
$a = 0.6$	0.232	0.138	0.100	1.952	1.615	1.446

Таблица 10. Верхние процентные точки распределения статистики критерия Андерсона-Дарлинга при проверке сложных гипотез о согласии с логарифмически нормальным законом

Степень цензурирования	Уровень значимости α					
	Цензурирование справа			Цензурирование слева		
	0.01	0.05	0.1	0.01	0.05	0.1
$a = 0$	3.278	2.022	1.51	3.278	2.022	1.51
$a = 0.1$	3.546	2.202	1.638	3.613	2.310	1.776
$a = 0.2$	3.262	2.001	1.48	3.278	2.022	1.51
$a = 0.3$	2.853	1.742	1.282	3.33	2.068	1.532
$a = 0.4$	2.377	1.435	1.053	5.199	3.575	2.856
$a = 0.5$	1.824	1.090	0.802	9.749	7.684	6.666
$a = 0.6$	1.244	0.748	0.552	17.245	14.927	13.737

На рисунках 4 и 5 представлены функции распределения статистики Колмогорова при проверке сложной гипотезы о согласии с законом Вейбулла при цензурировании II типа справа и слева соответственно. Из рисунков следует, что при увеличении степени цензурирования функции распределений по цензурированным справа данным смещаются влево, по цензурированным слева данным – вправо.

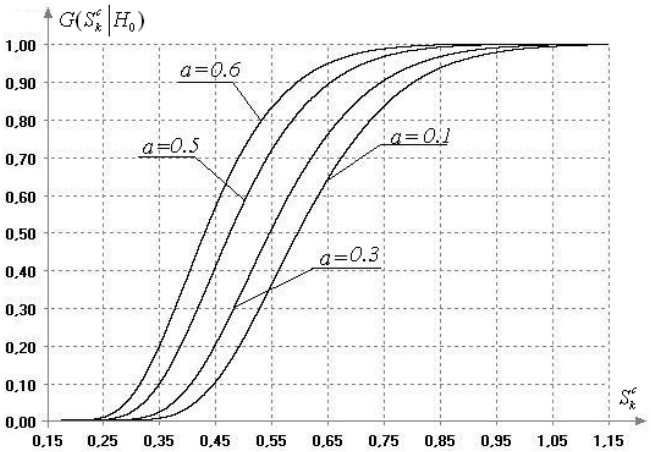

Рис. 4. Функции распределения статистики Колмогорова при проверке сложной гипотезы о согласии с законом Вейбулла при цензурировании II типа справа

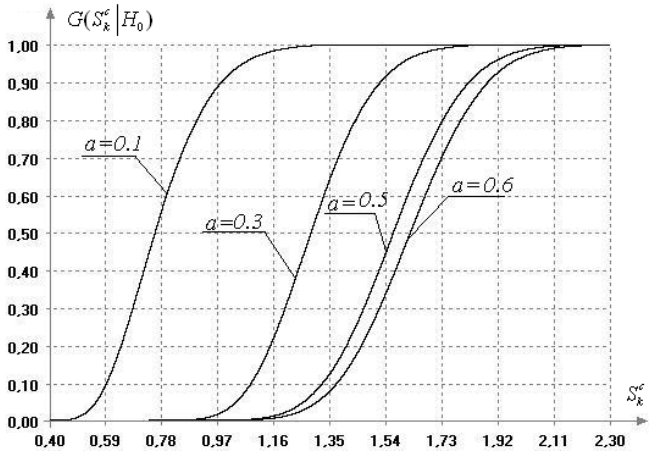

Рис. 5. Функции распределения статистики Колмогорова при проверке сложной гипотезы о согласии с законом Вейбулла при цензурировании II типа слева

6. ИССЛЕДОВАНИЕ МОЩНОСТИ НЕПАРАМЕТРИЧЕСКИХ КРИТЕРИЕВ

В [14, 15] приведены результаты сравнительного анализа мощности критериев согласия по полным выборкам. Однако следует ожидать, что в случае цензурированных наблюдений картина с мощностью критериев будет несколько иной. В настоящей работе мощность критериев согласия при различных парах H_0 и H_1 близких конкурирующих гипотез исследовалась в зависимости от степени цензурирования. Рассмотрены ситуации проверки простых и сложных гипотез.

При проверке простых гипотез в качестве примера рассмотрено три пары конкурирующих гипотез.

1. H_0 : распределение Вейбулла с плотностью

$$f(x)=\frac{\theta_0(x-\theta_2)^{\theta_0-1}}{\theta_1^{\theta_0}}\exp\left\{-\left(\frac{x-\theta_2}{\theta_1}\right)^{\theta_0}\right\}$$

и параметрами $\theta_0=2$, $\theta_1=2$, $\theta_2=0$; H_1 – гамма-распределение с плотностью

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

и параметрами $\theta_0=3.1215$, $\theta_1=0.5577$, $\theta_2=0$, при которых гамма-распределение наиболее близко к данному распределению Вейбулла (см. рис. 6).

2. H_0 : нормальное распределение с плотностью

$$f(x)=\frac{1}{\theta_0\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta_1)^2}{2\theta_0^2}\right\},$$

H_1 – логистическое распределение с функцией плотности

$$f(x)=\frac{\pi}{\theta_0\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\right]^2$$

и параметрами $\theta_0=2$, $\theta_1=0$. Эти два закона близки и трудно различимы с помощью критериев согласия (см. рис.7).

3. H_0: нормальное распределение с параметрами $\theta_0 = 2$, $\theta_1 = 0$, H_1: нормальное распределение с параметрами $\theta_0 = 2$, $\theta_1 = 0.15$ (см. рис. 8).

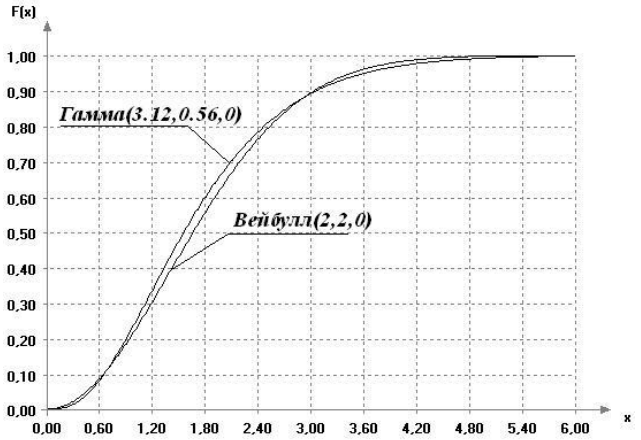

Рис. 6. Функции гамма-распределения и распределения Вейбулла

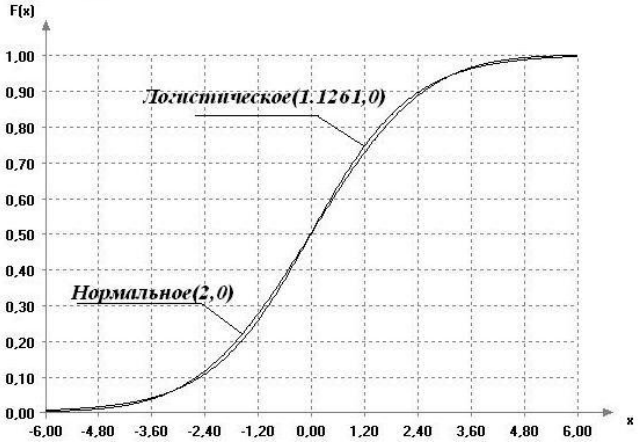

Рис. 7. Функции распределений нормального и логистического

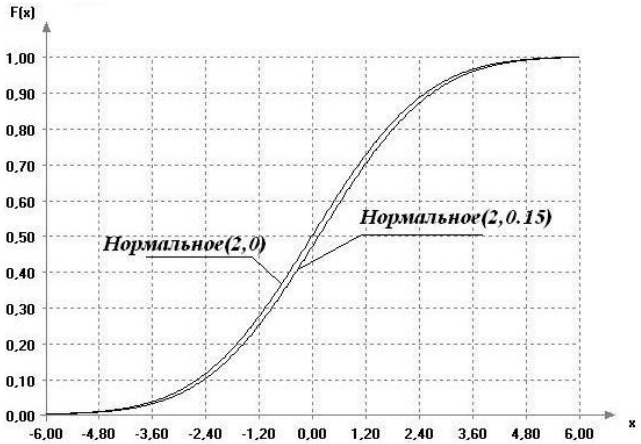

Рис. 8. Функции нормальных распределений со сдвигом 0.15

На рис. 9 представлены оценки мощности критериев Колмогорова, Андерсона-Дарлинга, Крамера-Мизеса-Смирнова в зависимости от степени цензурирования a, полученные при полных объемах выборок $n = 300$ и уровне значимости $\alpha = 0.1$, по *цензурированным справа* выборкам II типа для первой пары конкурирующих гипотез, законов распределения гамма- и Вейбулла. На рисунках 10-11 для второй и третьей пар конкурирующих гипотез, соответственно.

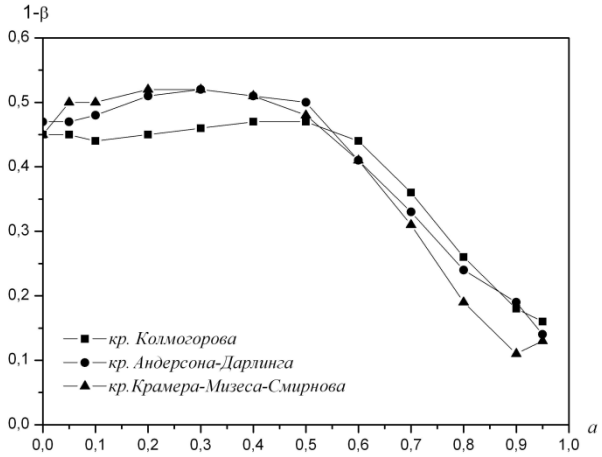

Рис. 9. Оценка мощности критериев в случае первой пары конкурирующих гипотез

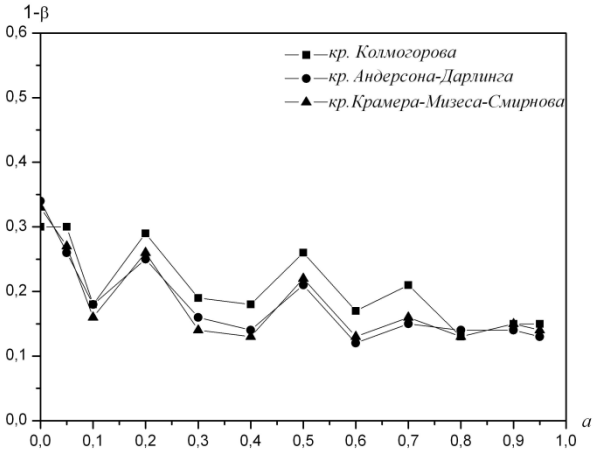

Рис. 10. Оценка мощности критериев в случае второй пары конкурирующих гипотез

Мощность критериев зависит от степени цензурирования, причем зависимость не является монотонной, и характер изменения мощности существенно зависит от вида проверяемых гипотез. Поведение функций мощности становится понятным, если проследить по рисункам 6-8 доступную для наблюдений $1-a$ часть области определения случайных величин. Например, хорошо объясняется, почему при увеличении степени цензурирования a от 0.1 до 0.5-0.6 в случае первой пары конкурирующих гипотез мощность критериев увеличивается (в наблюдаемой области различие законов оказывается более заметным, см. рис. 6). И наоборот, другая изменчивая картина поведения мощности в зависимости от a для второй пары конкурирующих гипотез, опять-таки, связанная с видом законов и их различием в наблюдаемой области (см. рис. 7).

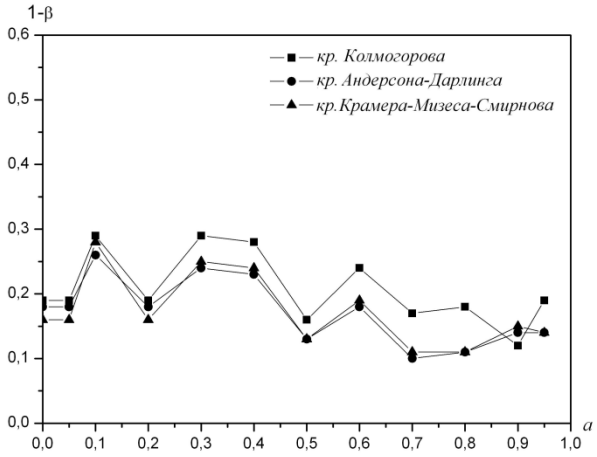

Рис. 11. Оценка мощности критериев в случае третьей пары конкурирующих гипотез

Из рисунков 9-11 видно, что при больших степенях цензурирования наибольшую мощность показал критерий Колмогорова, в то время как в случае полных выборок наиболее мощным чаще всего оказывается критерий Андерсона-Дарлинга [9, 10].

Характер изменения мощности при разных степенях цензурирования оказывается аналогичным в случае цензурирования слева и в случае цензурирования I типа.

При проверке сложных гипотез в качестве примера исследовалась мощность относительно первых двух пар рассмотренных выше конкурирующих гипотез: распределение Вейбулла – гамма-распределение и нормальное распределение – логистическое распределение.

На рис. 12 и 13 представлены оценки мощности критериев Колмогорова, Андерсона-Дарлинга, Крамера-Мизеса-Смирнова в зависимости от степени цензурирования a, полученные при полных объемах выборок $n = 300$ и уровне значимости $\alpha = 0.1$, по *цензурированным справа* выборкам II типа для первой и второй пар конкурирующих гипотез, соответственно. Как и в случае проверки простых гипотез, характер изменения мощности зависит от вида конкурирующих законов.

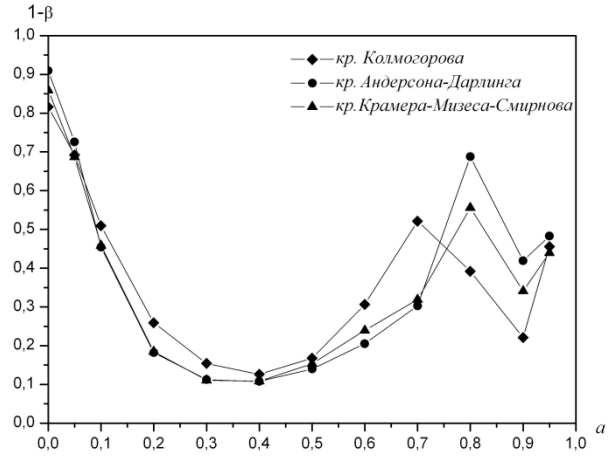

Рис. 12. Оценка мощности критериев в случае первой пары конкурирующих гипотез

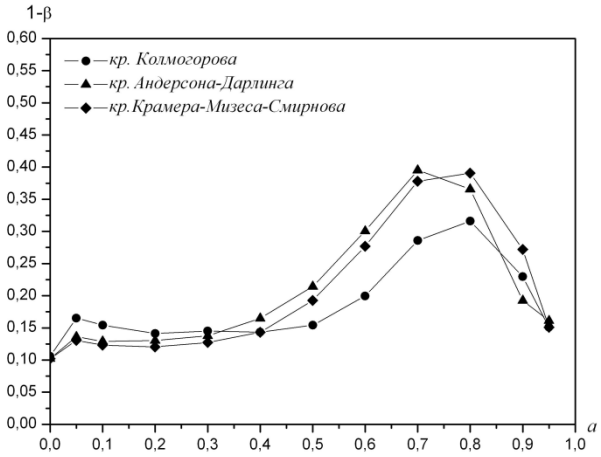

Рис. 13. Оценка мощности критериев в случае второй пары конкурирующих гипотез

Из рисунков 12 и 13 видно, что при высокой степени цензурирования $a = 0.7 - 0.8$ на графиках отмечается рост оценок мощности. К сожалению, этот, казалось бы, положительный факт объясняется ростом смещения оценок (и потерей информации), в связи с чем получаемая модель закона в большей степени отличается от истинной. Смещенность оценок оказывается самым сложным препятствием на пути решения проблемы проверки сложных гипотез по цензурированным данным о согласии наблюдаемого эмпирического распределения с теоретическим (при использовании любых критериев согласия).

В случае проверки сложных гипотез предпочтительность какого-либо критерия неочевидна, так как, обладая большей мощностью при одних степенях цензурирования, критерий может проигрывать при других. Последнее, опять же, объяснимо, так как непараметрические критерии по-разному улавливают отклонения на “хвостах” и в центре области определения случайных величин.

7. ЗАКЛЮЧЕНИЕ

Наиболее весомые результаты, полученные в классической математической статистике, имеют асимптотический характер. Однако на практике всегда имеют дело с ограниченными объемами наблюдений, а в такой ситуации свойства и распределения статистик могут существенно отличаться от асимптотических, что еще раз подчеркивает данная работа.

В результате проведенных исследований показано, что при ограниченных объемах выборок распределения статистики Реньи существенно зависят от степени цензурирования. При высокой или, наоборот, при малой степени цензурирования предельным распределением $L(S)$ можно пользоваться лишь при потенциальном объеме выборки $n > 1000$, что на практике редко выполнимо. Поэтому использовать критерий типа Реньи в практике статистического анализа цензурированных выборок не рекомендуется.

Распределения статистики типа Колмогорова при проверке простых гипотез быстро сходятся к соответствующему предельному закону $K_a^c(S)$. Для различной величины степени цензурирования найдены величины объема выборок, начиная с которых обеспечивается корректное применение критерия типа Колмогорова.

Методами компьютерного моделирования проведено исследование распределений статистик типа Колмогорова, Андерсона-Дарлинга, Крамера-Мизеса-Смирнова при проверке простых и сложных гипотез. Построены таблицы верхних процентных

точек для распределений статистик рассматриваемых непараметрических критериев согласия при проверке простых гипотез. Аналогичные таблицы построены для проверки сложных гипотез о согласии с законами Вейбулла и логарифмически нормальным.

Показано, что при различных степенях цензурирования мощность критериев согласия существенно зависит от вида проверяемых гипотез, при этом характер изменения мощности при разных степенях цензурирования не зависит от вида и типа цензурирования.

Исследования выполнены при финансовой поддержке Министерства образования и науки Российской Федерации в рамках федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009-2013 год».

Список литературы

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

[2] Вероятность и математическая статистика: Энциклопедия. Под ред. **Прохорова Ю.В.** – М., Большая Российская энциклопедия, 1999. – 910 с.

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

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

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

[6] **Kac, M., Kiefer, J., Wolfowitz, J.** On tests of normality and other tests of goodness of fit based on distance methods // Annals of Mathematical Statistics. – 1955. V.26. – P.189–211.

[7] **Лемешко Б.Ю., Лемешко С.Б.** Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Часть 1 // Измерительная техника. – 2009. № 6. – С.6-11.

[8] **Лемешко Б.Ю., Лемешко С.Б.** Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч.II // Измерительная техника. – 2009. № 8. – С.17-26.

[9] **Lemeshko B.Yu. Lemeshko S.B. and Postovalov S.N.** Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses // Communications in Statistics - Theory and Methods, 2010. – Vol. 39, – No. 3. – P. 460-471.

[10] **Лемешко Б.Ю., Постовалов С.Н., Чимитова Е.В.** К оцениванию параметров законов распределений и проверке гипотез по цензурированным выборкам // Труды V международной конференции “Актуальные проблемы электронного приборостроения” АПЭП-2000. – Т.7, Новосибирск, 2000. – С.188-191.

[11] **Лемешко Б.Ю., Чимитова Е.В.** К проверке простых гипотез о согласии по дискретным, группированным или цензурированным данным с использованием непараметрических критериев // Материалы конференции ASMDA, Греция, Крит, 2007.

[12] **Лемешко Б.Ю.** Об оценивании параметров распределений и проверке гипотез по цензурированным выборкам // Методы менеджмента качества. 2001. – № 4. – С.32-38.

[13] **Лемешко Б.Ю., Гильдебрант С.Я., Постовалов С.Н.** К оцениванию параметров надежности по цензурированным выборкам // Заводская лаборатория. Диагностика материалов. 2001. – Т.67. – № 1. – С. 52-64.

[14] **Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н.** Сравнительный анализ мощности критериев согласия при близких конкурирующих гипотезах. I. Проверка простых гипотез // Сибирский журнал индустриальной математики. 2008. – Т.11. – № 2(34). – С.96-111.

[15] **Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н.** Сравнительный анализ мощности критериев согласия при близких альтернативах. II. Проверка сложных гипотез // Сибирский журнал индустриальной математики. 2008. – Т.11. – № 4(36). – С.78-93.

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 4 монографии.

Чимитова Екатерина Владимировна, кандидат технических наук, доцент кафедры прикладной математики НГТУ. Основное направление научных исследований – статистические методы теории надежности и анализа выживаемости. Имеет более 40 публикаций.

Плешкова Татьяна Александровна, бакалавр, магистрант второго года. Основное направление научных исследований – статистические методы анализа цензурированных данных.

Testing simple and composite goodness-of-fit hypotheses by censored samples. Lemeshko B.Yu., Chimitova E.V., Pleshkova T.A.

The problems of application of nonparametric Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling goodness-of-fit tests for censored data have been considered in this paper. The convergence of statistic distributions to the corresponding limiting distribution laws has been investigated under true null hypothesis by means of statistical simulation methods, as well as the test power against close competing hypotheses. The distributions of considered test statistics have been investigated for composite hypotheses.

Keywords: goodness-of-fit tests, censored data, Renyi test, Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test.