Сибирский государственный университет
телекоммуникаций и информатики

ОБРАБОТКА ИНФОРМАЦИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

РОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ

МАТЕРИАЛЫ КОНФЕРЕНЦИИ

Новосибирск
2018

ISBN 978-5-91434-042-8

СОДЕРЖАНИЕ

Секция 1
ИНФОРМАТИКА И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
Подсекция 1.1. НГТУ

УДК 519.24

О применении и мощности k-выборочных критериев однородности законов

Б. Ю. Лемешко, И. В. Веретельникова[1]

Новосибирский государственный технический университет

Исследованы свойства k-выборочных критериев однородности законов распределения. Предложены критерии, в качестве статистик которых используется максимум 2-выборочных статистик критериев Смирнова, Лемана–Розенблатта и Андерсона–Дарлинга, применяемых к попарно сравниваемым k выборкам. Построены модели предельных распределений статистик для предложенных критериев, а также для k-выборочного критерия Андерсона–Дарлинга. Проведен сравнительный анализ мощности критериев.

Ключевые слова: k-выборочные критерии, критерий однородности, мощность критерия, статистическое моделирование

1. Введение

С необходимостью решения задач проверки гипотез о принадлежности двух (или более) выборок случайных величин одной и той же генеральной совокупности (проверки однородности) сталкиваются в различных областях. Например, такая задача естественно возникает при проверке средств измерений, когда пытаются убедиться в том, что закон распределения случайных ошибок измерений не претерпел существенных изменений по истечении некоторого интервала времени.

Задача проверки однородности k выборок формулируется следующим образом. Пусть x_{ij} j-е наблюдение в вариационном ряду i-й выборки $j=\overline{1,n_i}$, $i=\overline{1,k}$. Предположим, что i-й выборке соответствует непрерывная функция распределения $F_i(x)$. Необходимо проверить гипотезу вида $H_0 : F_1(x) = F_2(x) = \ldots = F_k(x)$ без указания общего закона распределения.

Как правило, на практике используется двухвыборочные критерии Смирнова [1] и Лемана–Розенблатта [1, 2, 3]. Предпочтительность использования данных критериев для проверки однородности обсуждалась в [4]. В русскоязычной литературе практически не упоминается о применении двухвыборочного критерия Андерсона–Дарлинга [5] (Андерсона–Дарлинга–Петита) или, тем более, об использовании k-выборочного варианта критерия Андерсона–Дарлинга [6] или о критериях Жанга [7, 8, 9].

Критерий однородности Смирнова предложен в работе [10]. Предполагается, что функции распределения $F_1(x)$ и $F_2(x)$ являются непрерывными. Статистика критерия Смирнова измеряет расстояние между эмпирическими функциями распределения, построенными по выборкам

$$D_{n_2,n_1} = \sup_x \left| F_{2n_1}(x) - F_{1n_2}(x) \right| .$$

[1] Работа выполнена при поддержке Министерства образования и науки РФ в рамках государственной работы «Обеспечение проведения научных исследований» (№ 1.4574.2017/6.7) и проектной части государственного задания (№ 1.1009.2017/4.6).

При практическом использовании критерия значение статистики D_{n_1,n_2} рекомендуется вычислять в соответствии с соотношениями [1]:

$$D^+_{n_2,n_1} = \max_{1\le r\le n_2}\left[\frac{r}{n_2} - F_{1n_1}(x_{2r})\right] = \max_{1\le s\le n_1}\left[F_{2n_2}(x_{2s}) - \frac{s-1}{n_1}\right],$$

$$D^-_{n_2,n_1} = \max_{1\le r\le n_2}\left[F_{1n_1}(x_{2r}) - \frac{r-1}{n_2}\right] = \max_{1\le s\le n_1}\left[\frac{s}{n_1} - F_{2n_2}(x_{1s})\right],$$

$$D_{n_2,n_1} = \max\left(D^+_{n_2,n_1}, D^-_{n_2,n_1}\right).$$

Если гипотеза H_0 справедлива, то при неограниченном увеличении объемов выборок статистика

$$S_C = \sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_2,n_1} \quad (1)$$

в пределе подчиняется распределению Колмогорова $K(S)$ [1] .

Статистика **критерия Лемана–Розенблатта**, предложенного в работе [2], используется в форме [1]

$$T = \frac{1}{(n_1+n_2)}\left[n_2\sum_{i=1}^{n_2}\left(r_i - i\right)^2 + n_1\sum_{j=1}^{n_1}\left(s_j - j\right)^2\right] - \frac{4n_1n_2-1}{6\ n_1+n_2()}, \quad (2)$$

где r_i – порядковый номер (ранг) x_{2i}; s_j – порядковый номер (ранг) x_{1j} в объединенном вариационном ряде. В [3] было показано, что статистика (2) в пределе распределена как $a1(t)$.

Двухвыборочный **критерий Андерсона–Дарлинга** рассмотрен в работе [5]. Статистика критерия определяется выражением

$$A^2 = \frac{1}{n_1n_2}\sum_{i=1}^{n_1+n_2-1}\frac{\left(M_i(n_1+n_2) - n_1 i\right)^2}{i(n_1+n_2-i)}, \quad (3)$$

где M_i – число элементов первой выборки, меньших или равных *i*-му элементу вариационного ряда объединенной выборки. Предельным распределением статистики (3) при справедливости проверяемой гипотезы H_0 является распределение $a2(t)$.

2. k-выборочный критерий Андерсона–Дарлинга

Вопросы построения *k*-выборочных критериев однородности законов, являющихся аналогами критериев согласия Колмогорова–Смирнова и Крамера–Мизеса (*k*-выборочными аналогами критериев однородности Смирнова и Лемана–Розенблатта), рассматривались в работе [11]. Многовыборочный вариант критерия согласия Андерсона–Дарлинга предложен в [6].

Проверяется гипотеза вида $H_0: F_1(x) = F_2(x) = \ldots = F_k(x)$. Пусть x_{ij} j-е наблюдение i-й выборки $j = \overline{1, n_i}$, $i = \overline{1, k}$, i-й выборке соответствует непрерывная функция распределения $F_i(x)$. Обозначим эмпирическую функцию распределения, соответствующую i-й выборке, как $F_{in_i}(x)$, а эмпирическую функцию распределения, соответствующую объединённой выборке объёмом $n = \sum_{i=1}^{k} n_i$, как $H_n(x)$. Статистика *k*-выборочного критерия Андерсона–Дарлинга определяется выражением

$$A_{kn}^2 = \sum_{i=1}^{k} n_i \int_{B_n} \frac{[F_{in_i}(x) - H_n(x)]^2}{\left(1 - H_n(x)\right) H_n(x)} dH_n(x),$$

где $B_n = \{x \in R:\ H_n(x) < 1\}$. В предположении о непрерывности $F_i(x)$, упорядочив объединённую выборку $Z_1 \le Z_2 \le \ldots \le Z_n$, можно получить простое выражение для вычисления статистики [6]:

$$A_{kn}^2 = \frac{1}{n} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n-1} \frac{\left(nM_{ij} - jn_i\right)^2}{j(n-j)},$$

где M_{ij} – число элементов в i-й выборке, которые не больше чем Z_j. Проверяемая гипотеза H_0 отклоняется при больших значениях статистики.

В работе [6] статистика приобретает следующий окончательный вид:

$$T_{kn} = \frac{A_{kn}^2 - (k-1)}{\sqrt{D[A_{kn}^2]}}, \qquad (4)$$

где дисперсия статистики A_{kn}^2 определяется выражением [6]

$$D\left[A_{kn}^2\right] = \frac{an^3 + bn^2 + cn + d}{(n-1)(n-2)(n-3)}$$

при

$$a = (4g-6)(k-1) + (10-6g)H,$$

$$b = (2g-4)k^2 + 8hk + (2g-14h-4)H - 8h + 4g - 6,$$

$$c = (6h+2g-2)k^2 + (4h-4g+6)k + (2h-6)H + 4h,$$

$$d = (2h+6)k^2 - 4hk,$$

где

$$H = \sum_{i=1}^{k} \frac{1}{n_i}, \quad h = \sum_{i=1}^{n-1} \frac{1}{i}, \quad g = \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \frac{1}{(n-i)j}.$$

Зависимость предельных распределений статистики (4) от числа сравниваемых выборок k иллюстрирует рис. 1. С ростом числа сравниваемых выборок это распределение медленно сходится к стандартному нормальному закону.

В [6] для статистики (4) для ряда k построена таблица критических значений. В [12, 13, 14] на основании результатов статистического моделирования нами построены модели предельных распределений статистики (4) для $k = 2 \div 11$. Хорошими моделями оказались законы семейства бета-распределений III рода с плотностью

$$f(x) = \frac{\theta_2^{\theta_0}}{\theta_3 \mathrm{B}(\theta_0, \theta_1)} \left(\frac{x-\theta_4}{\theta_3}\right)^{\theta_0 - 1} \left(1 - \frac{x-\theta_4}{\theta_3}\right)^{\theta_1 - 1} \Bigg/ \left[1 + (\theta_2 - 1)\frac{x-\theta_4}{\theta_3}\right]^{\theta_0 + \theta_1} \qquad (5)$$

при конкретных значениях параметров этого закона $\mathrm{B}_{\mathrm{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$, найденными по выборкам статистик объёмом $N = 10^6$, полученным в результате моделирования.

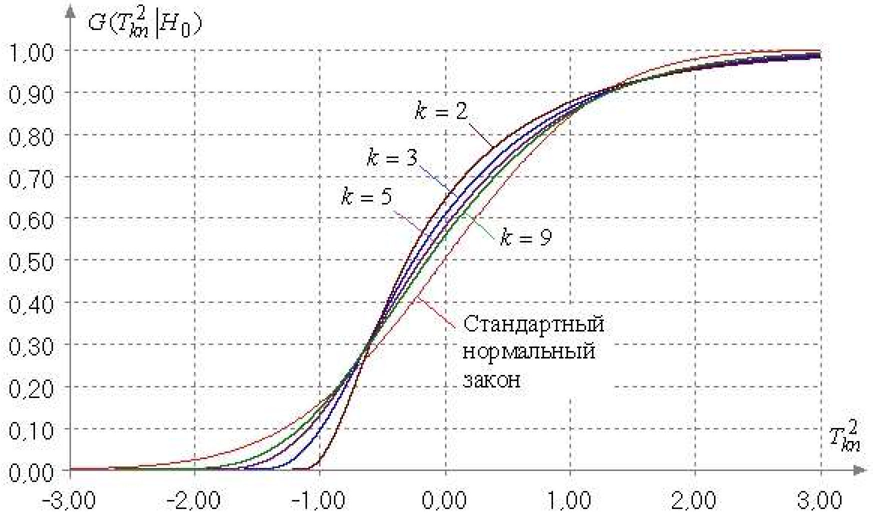

Рис. 1. Зависимость предельных распределений статистики (4) от числа сравниваемых выборок

Представленные в таблице 1 модели $\mathrm{B_{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$, с приведенными значениями параметров, позволяют по значениям статистики, вычисленным в соответствии с соотношением (4), находить оценки p_{value} при соответствующем числе k сравниваемых выборок.

Таблица 1. Модели предельных распределений статистики (4)

k	Модель
2	B_{III} (3.1575, 2.8730, 18.1238, 15.0000, –1.1600)
3	B_{III} (3.5907, 4.5984, 7.8040, 14.1310, –1.5000)
4	B_{III} (4.2657, 5.7035, 5.3533, 12.8243, –1.7500)
5	B_{III} (6.2992, 6.5558, 5.6833, 13.010, –2.0640)
6	B_{III} (6.7446, 7.1047, 5.0450, 12.8562, –2.2000)
7	B_{III} (6.7615, 7.4823, 4.0083, 11.800, –2.3150)
8	B_{III} (5.8057, 7.8755, 2.9244, 10.900, –2.3100)
9	B_{III} (9.0736, 7.4112, 4.1072, 10.800, –2.6310)
10	B_{III} (10.2571, 7.9758, 4.1383, 11.186, –2.7988)
11	B_{III} (10.6848, 7.5950, 4.2041, 10.734, –2.8400)
∞	N(0.0, 1.0)

3. Критерии Жанга

Предложенные Жангом критерии [7, 8, 9] являются развитием критериев однородности Смирнова, Лемана–Розенблатта и Андерсона–Дарлинга, они дают возможность сравнивать $k \geq 2$ выборок.

Пусть $x_{i1}, x_{i2}, \ldots, x_{in_i}$ упорядоченные выборки непрерывных случайных величин с функциями распределения $F_i(x)$, $(i = \overline{1,k})$ и пусть $X_1 < X_2 < \ldots < X_n$, где $n = \sum_{i=1}^{k} n_i$, объединённая упорядоченная выборка.

Обозначим R_{ij} ранг j-го упорядоченного наблюдения x_{ij} i-й выборки в объединённой выборке. Пусть $X_0 = -\infty$, $X_{n+1} = +\infty$, а ранги $R_{i,0} = 1$, $R_{i,n_i+1} = n+1$.

В критериях используется модификация эмпирической функции распределения $\hat{F}(t)$, принимающая в точках разрыва X_m, $m=\overline{1,n}$, значения $\hat{F}(X_m)=(m-0.5)/n$ [7].

Статистика Z_K критерия однородности Жанга имеет вид [7]:

$$Z_K = \max_{1\le m\le n}\left\{\sum_{i=1}^{k} n_i\left[F_{i,m}\ln\frac{F_{i,m}}{F_m}+\left(1-F_{i,m}\right)\ln\frac{1-F_{i,m}}{1-F_m}\right]\right\}, \tag{6}$$

где $F_m=\hat{F}(X_m)$, так что $F_m=(m-0.5)/n$, а вычисление $F_{i,m}=\hat{F}_i(X_m)$ осуществляется следующим образом. В начальный момент значения $j_i=0$, $i=\overline{1,k}$. Если $R_{i,j_i+1}=m$, то $j_i := j_i+1$ и $F_{i,m}=(j_i-0.5)/n_i$, в противном случае если $R_{i,j_i}<m<R_{i,j_i+1}$, то $F_{i,m}=j_i/n_i$.

Критерий *правосторонний*: проверяемая гипотеза H_0 отклоняется при *больших* значениях статистики (6). Зависимость распределений статистик от объема выборок и их числа демонстрируется на рисунках 2 и 3.

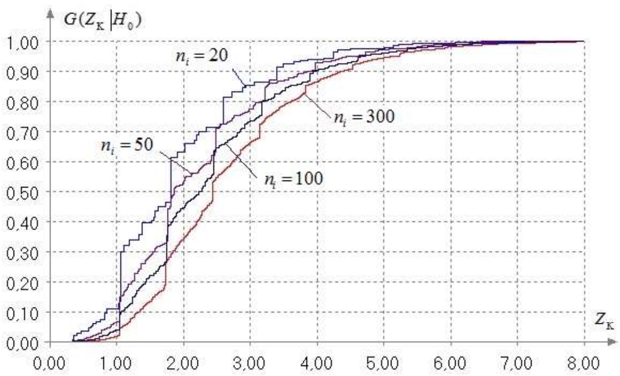

Рис. 2. Зависимость распределений статистики (6) от объёмов выборок

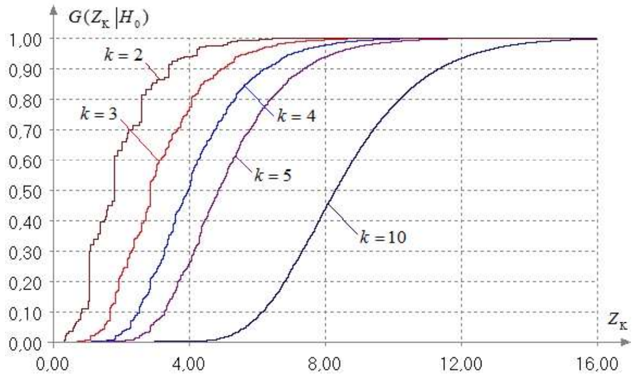

Рис. 3. Зависимость распределений статистики (6) от k при $n_i=20$

Статистика Z_A критерия однородности k выборок определяется выражением [7]:

$$Z_A = -\sum_{m=1}^{n}\sum_{i=1}^{k} n_i \frac{F_{i,m}\ln F_{i,m}+\left(1-F_{i,m}\right)\ln\left(1-F_{i,m}\right)}{(m-0.5)(n-m+0.5)}, \tag{7}$$

где F_m и $F_{i,m}$ вычисляются, как определено выше.

Критерий *левосторонний*: проверяемая гипотеза H_0 отклоняется при *малых* значениях статистики (7). Зависимость распределений статистик от объема выборок и их числа демонстрируется на рисунках 4 и 5.

Статистика Z_C критерия однородности k выборок определяется выражением [7]:

$$Z_C = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln\left(\frac{n_i}{j-0.5}-1\right)\ln\left(\frac{n}{R_{i,j}-0.5}-1\right). \tag{8}$$

Этот критерий также *левосторонний*: проверяемая гипотеза H_0 отклоняется при *малых* значениях статистики (8). Распределения статистики (8) также зависят от объема выборок и числа сравниваемых выборок.

Зависимость распределений статистик (6) – (8) от объёмов выборок осложняет использование критериев Жанга, так как возникают проблемы с вычислением оценки p_{value}.

В тоже время, отсутствие информации о законах распределения статистик и таблиц критических значений в современных условиях не является серьёзным недостатком критериев Жанга, так как в программном обеспечении, осуществляющем поддержку применения критериев, несложно организовать вычисление достигнутых уровней значимостей p_{value}, используя методы статистического моделирования.

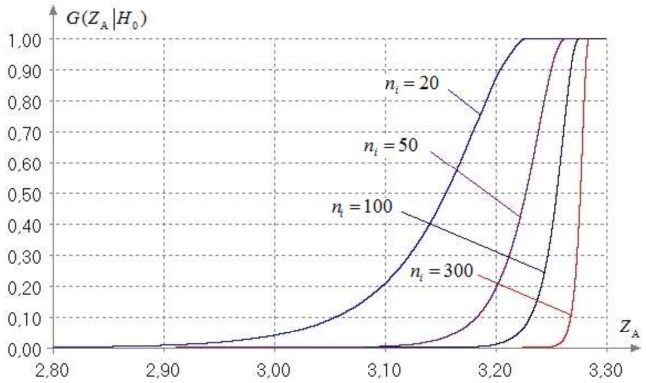

Рис.4. Зависимость распределений статистики (7) от объёмов выборок

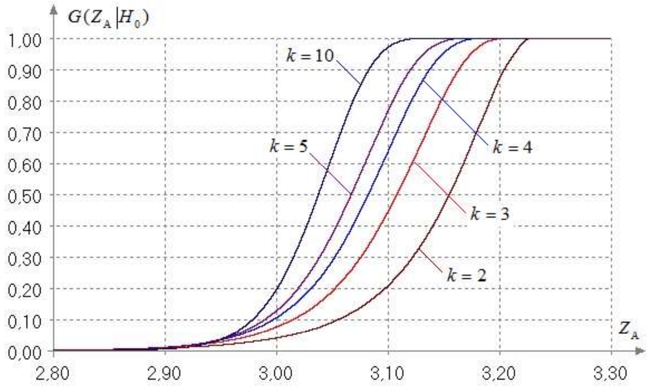

Рис. 5. Зависимость распределений статистики (7) от k при $n_i = 20$

4. Анализ k выборок с использованием двухвыборочных критериев

Различные подходы к построению k-выборочных аналогов критериев Смирнова, Лемана–Розенблатта и Андерсона–Дарлинга рассматривались в работе [11]. k-выборочный вариант критерия Колмогорова–Смирнова, основанный на таком подходе, был построен в [15] и рассматривается в последовательных изданиях книги [16]. На таком же подходе построен k-выборочный критерий Андерсона–Дарлинга [6], рассмотренный. В этих критериях, так же как и в критериях однородности Жанга, строится объединённая выборка, а статистики измеряют отклонение эмпирических распределений отдельных выборок от эмпирического распределения, построенного по совокупности анализируемых выборок.

Возможен другой путь. Для анализа k выборок можно к каждой паре выборок применить двухвыборочный критерий со статистикой S (всего $(k-1)k/2$ вариантов), а решение принимать по совокупности результатов. В качестве статистики такого k-выборочного критерия (в случае использования правостороннего двухвыборочного критерия) можно рассмотреть, например, статистику вида

$$S_{\max} = \max_{\substack{1\le i\le k \\ i<j\le k}} \{S_{i,j}\}, \tag{9}$$

где $S_{i,j}$ – значения статистик используемого двухвыборочного критерия, полученные при анализе i-й и j-й выборок.

Проверяемая гипотеза H_0 будет отклоняться при **больших** значениях статистики $S_{\max}$. Преимуществом такого рода критерия является и то, что в результате будет определена пара выборок, различие между которыми оказывается наиболее значимым с позиций используемого двухвыборочного критерия.

В качестве $S_{i,j}$ можно использовать статистики двухвыборочных критериев Смирнова (лучше в модифицированном виде [17]), Лемана–Розенблата, Андерсона–Дарлинга. В этом случае распределения соответствующих статистик $S_{\max}$ сходятся к некоторым предельным, модели которых могут быть найдены по результатам статистического моделирования.

В случае **k-выборочного варианта критерия Смирнова** в качестве $S_{i,j}$ в (9) будет рассматриваться модификация статистики Смирнова [17]

$$S_{mod} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\left(D_{n_2,n_1} + \frac{n_1+n_2}{4.6 n_1 n_2}\right), \tag{10}$$

распределение которой всегда ближе к предельному распределению Колмогорова $K(S)$. Статистику $S_{\max}$ в этом случае будем обозначать как $S_{\max}^{Sm}$.

При равных объёмах сравниваемых выборок распределения статистики $S_{\max}^{Sm}$ (как и в двухвыборочном варианте) обладают существенной дискретностью (см. рис. 6) и отличаются от асимптотических (предельных) распределений (см. рис. 7). Если есть такая возможность, то предпочтительней в качестве n_i выбирать взаимно простые числа, тогда распределения $G(S|H_0)$ статистики $S_{\max}^{Sm}$ практически не будут отличаться от асимптотических.

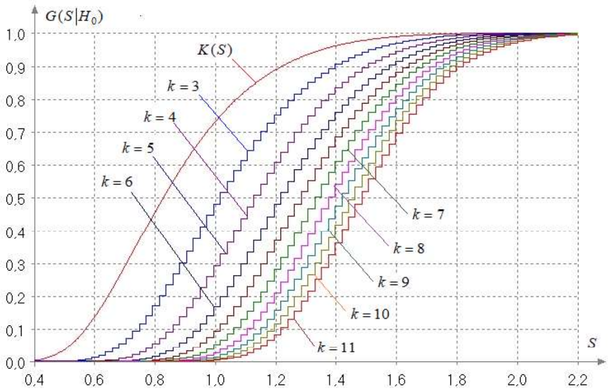

Рис. 6. Распределения статистики $S_{\max}^{Sm}$, $n_i = 1000$

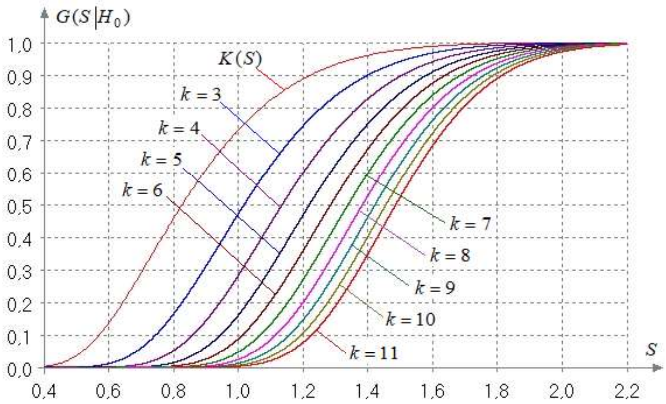

Рис. 7. Асимптотические распределения статистики $S_{\max}^{Sm}$

Построенные по эмпирическим распределениям статистик, полученным методом Монте–Карло при количестве имитационных экспериментов $N = 10^6$, модели асимптотических (предельных) распределений статистики $S_{\max}^{Sm}$ при числе сравниваемых выборок $k = 3 \div 11$ представлены в таблице 2.

Таблица 2. Модели предельных распределений статистики $S_{\max}^{Sm}$

k	Модель
2	$K(S)$
3	B_{III} (6.3274, 6.6162, 2.8238, 2.4073, 0.4100)
4	B_{III} (7.2729, 7.2061, 2.6170, 2.3775, 0.4740)
5	B_{III} (7.1318, 7.3365, 2.4813, 2.3353, 0.5630)
6	B_{III} (7.0755, 8.0449, 2.3163, 2.3818, 0.6320)
7	B_{III} (7.7347, 8.6845, 2.3492, 2.4479, 0.6675)
8	B_{III} (7.8162, 8.9073, 2.2688, 2.4161, 0.7120)
9	B_{III} (7.8436, 8.8805, 2.1696, 2.3309, 0.7500)
10	B_{III} (7.8756, 8.9051, 2.1977, 2.3280, 0.7900)
11	B_{III} (7.9122, 9.0411, 2.1173, 2.2860, 0.8200)

Представленные в таблице 2 модели $\mathrm{B}_{\mathrm{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ бета-распределения 3-го рода (5) с приведенными значениями параметров, позволяют по значениям статистики, вычисленным в соответствии с соотношением (9) с использованием в качестве $S_{i,j}$ статистики Смирнова (1) или её модификации (10), находить оценки p_{value} при соответствующем числе k сравниваемых выборок.

В случае k**-выборочного варианта критерия Лемана–Розенблатта** в качестве $S_{i,j}$ в статистике $S_{\max}^{LR}$ вида (9) используется статистика (2). Зависимость распределений статистики при справедливости H_0 от числа выборок иллюстрирует рис. 8.

Построенные модели асимптотических (предельных) распределений статистики $S_{\max}^{LR}$ при числе сравниваемых выборок $k = 3 \div 11$ представлены в таблице 3. В данном случае наилучшими моделями оказались распределения Sb–Джонсона с плотностью

$$f(x) = \frac{\theta_1 \theta_2}{\sqrt{2\pi}(x-\theta_3)(\theta_2+\theta_3-x)} \exp\left\{ -\frac{1}{2}\left[\theta_0 - \theta_1 \ln \frac{x-\theta_3}{\theta_2+\theta_3-x}\right]^2 \right\}$$

при конкретных значениях параметров этого закона, обозначенного в таблице 3 как $Sb(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$. Представленные модели позволяют по значениям статистики $S_{\max}^{LR}$ при соответствующем числе k сравниваемых выборок находить оценки p_{value}.

Таблица 3. Модели предельных распределений статистики $S_{\max}^{LR}$

k	Модель
2	$a1(t)$
3	Sb(3.2854, 1.2036, 3.0000, 0.0215)
4	Sb(2.5801, 1.2167, 2.2367, 0.0356)
5	Sb(3.1719, 1.4134, 3.1500, 0.0320)
6	Sb(2.9979, 1.4768, 2.9850, 0.0380)
7	Sb(3.2030, 1.5526, 3.4050, 0.0450)
8	Sb(3.2671, 1.6302, 3.5522, 0.0470)
9	Sb(3.4548, 1.7127, 3.8800, 0.0490)
10	Sb(3.4887, 1.7729, 3.9680, 0.0510)
11	Sb(3.4627, 1.8168, 3.9680, 0.0544)

В случае k**-выборочного варианта критерия Андерсона–Дарлинга** в качестве $S_{i,j}$ в статистике $S_{\max}^{AD}$ вида (9) используется статистика (3). Зависимость распределений статистики при справедливости H_0 от числа выборок иллюстрирует рис. 9.

Для распределений $G(S_{\max}^{AD} | H_0)$ также построены модели асимптотических (предельных) распределений статистики $S_{\max}^{AD}$ для числа сравниваемых выборок $k = 3 \div 11$, которые представлены в таблице 4. В этом случае лучшими моделями оказались бета-распределения 3-го рода (5), которые в виде $\mathrm{B}_{\mathrm{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ с конкретными значениями параметров приведены в таблице 4 и могут использоваться для оценки p_{value} при k сравниваемых выборках.

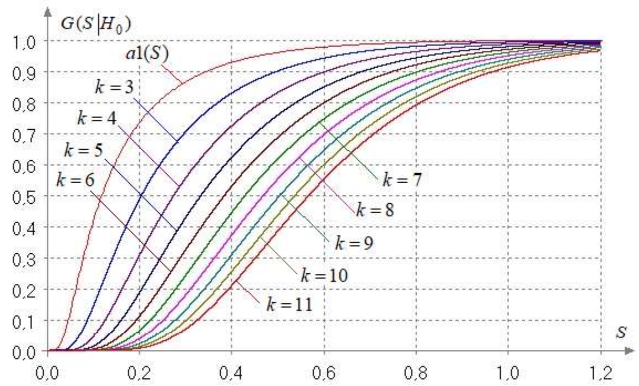

Рис. 8. Распределения статистики $S_{\max}^{LR}$

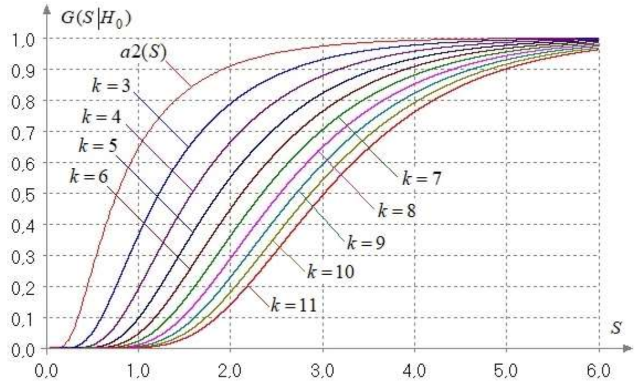

Рис. 9. Распределения статистики $S_{\max}^{AD}$

Таблица 4. Модели предельных распределений статистики $S_{\max}^{AD}$

k	Модель
2	$a2(t)$
3	B_{III} (4.4325, 2.7425, 12.1134, 8.500, 0.1850)
4	B_{III} (5.2036, 3.2160, 10.7792, 10.000, 0.2320)
5	B_{III} (5.7527, 3.3017, 9.7365, 10.000, 0.3000)
6	B_{III} (5.5739, 3.4939, 7.7710, 10.000, 0.3750)
7	B_{III} (6.4892, 3.6656, 8.0529, 10.500, 0.3920)
8	B_{III} (6.3877, 3.8143, 7.3602, 10.800, 0.4800)
9	B_{III} (6.7910, 3.9858, 7.1280, 11.100, 0.5150)
10	B_{III} (6.7533, 4.2779, 6.6457, 11.700, 0.5800)
11	B_{III} (7.1745, 4.3469, 6.6161, 11.800, 0.6100)

5. Сравнительный анализ мощности критериев

Одной из основных характеристик статистического критерия является его мощность относительно заданной конкурирующей гипотезы H_1, которая представляет собой разность $1-\beta$, где β – вероятность ошибки 2-го рода (принять гипотезу H_0 при справедливости H_1) при заданной вероятности α ошибки 1-го рода (отклонить H_0 при её справедливости).

Мощность рассматриваемых критериев однородности исследовалась методами статистического моделирования относительно трёх видов альтернатив: изменения параметра сдвига, изменения масштаба и относительно ситуации, когда пара выборок принадлежала близким, но различным законам (нормальному и логистическому).

Мощность k-выборочных критериев исследовалась для ситуаций, когда в качестве проверяемой гипотезы H_0 рассматривалась принадлежность всех выборок стандартному нормальному закону, в качестве конкурирующей гипотезы H_1 – принадлежность всех выборок, кроме последней, стандартному нормальному закону, а последней – нормальному закону с параметром сдвига $\theta_0 = 0.1$ и параметром масштаба $\theta_1 = 1$, в качестве гипотезы H_2 – принадлежность последней выборки нормальному закону с параметром сдвига $\theta_0 = 0$ и параметром масштаба $\theta_1 = 1.1$, в качестве конкурирующей гипотезы H_3 – принадлежность последней выборки логистическому закону с плотностью

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

и параметрами $\theta_0=0$ и $\theta_1=1$. Рассматривалась ситуация с анализом выборок одинакового объёма.

Оценки мощности находились по результатам моделирования распределений статистик при справедливости проверяемой $G(S|H_0)$ и конкурирующих гипотез $G(S|H_1)$, $G(S|H_2)$ и $G(S|H_3)$ при равных объёмах n_i сравниваемых выборок. В качестве примера в таблице 5 приведены оценки мощности критериев при $\alpha=0.1$ для $k=4$.

Таблица 5. Оценки мощности относительно альтернатив H_1, H_2 и H_3 ($k=4$, $n_i=n$)

Критерий	n =20	n =50	n = 100	n = 300	n = 500	n = 1000
Относительно альтернативы H_1						
$S_{\max}^{AD}$	0.112	0.131	0.165	0.302	0.438	0.706
AD	0.112	0.131	0.164	0.301	0.433	0.701
$S_{\max}^{LR}$	0.113	0.130	0.162	0.293	0.425	0.686
$S_{\max}^{Sm}$	0.111	0.125	0.151	0.261	0.366	0.605
Z_C	0.111	0.126	0.155	0.260	0.368	0.595
Z_{A}	0.111	0.127	0.153	0.255	0.360	0.579
Z_{K}	0.109	0.121	0.141	0.219	0.300	0.502
Относительно альтернативы H_2						
Z_C	0.106	0.122	0.158	0.306	0.468	0.761
Z_{A}	0.107	0.124	0.158	0.305	0.463	0.745
Z_{K}	0.106	0.120	0.145	0.249	0.367	0.606
AD	0.104	0.110	0.123	0.180	0.254	0.474
$S_{\max}^{AD}$	0.101	0.104	0.111	0.145	0.195	0.381
$S_{\max}^{Sm}$	0.102	0.105	0.108	0.128	0.153	0.221
$S_{\max}^{LR}$	0.102	0.103	0.105	0.118	0.135	0.197
Относительно альтернативы H_3						
Z_{A}	0.103	0.107	0.116	0.179	0.274	0.566
Z_C	0.103	0.107	0.115	0.173	0.257	0.555
Z_{K}	0.103	0.107	0.114	0.161	0.222	0.410
AD	0.102	0.106	0.113	0.143	0.179	0.291
$S_{\max}^{Sm}$	0.103	0.104	0.112	0.138	0.166	0.257
$S_{\max}^{AD}$	0.101	0.103	0.107	0.124	0.147	0.229
$S_{\max}^{LR}$	0.102	0.102	0.105	0.116	0.130	0.183

Анализ полученных оценок мощности позволяет сделать определённые выводы. Относительно конкурирующих гипотез, соответствующих изменению параметра сдвига, критерии можно упорядочить по мощности следующим образом:

$S_{\max}^{AD}$ $\succ$ Андерсона–Дарлинга $\succ$ $S_{\max}^{LR}$ $\succ$ $S_{\max}^{Sm}$ $\succ$ Жанга Z_C $\succ$ Жанга Z_{A} $\succ$ Жанга Z_{K}.

Относительно изменения параметра масштаба –

$$\text{Жанга } Z_C \succ \text{Жанга } Z_\text{А} \succ \text{Жанга } Z_\text{К} \succ \text{Андерсона–Дарлинга} \succ S_{\max}^{AD} \succ S_{\max}^{Sm} \succ S_{\max}^{LR}.$$

При этом критерии Жанга со статистиками $Z_\text{А}$ и Z_C практически эквивалентны по мощности, а критерий Андерсона–Дарлинга заметно уступает критериям Жанга.

Относительно ситуации, когда три выборки принадлежат нормальному закону, а четвёртая – логистическому, критерии располагаются по мощности в следующем порядке:

$$\text{Жанга } Z_\text{А} \succ \text{Жанга } Z_C \succ \text{Жанга } Z_\text{К} \succ \text{Андерсона–Дарлинга} \succ S_{\max}^{Sm} \succ S_{\max}^{AD} \succ S_{\max}^{LR}.$$

Можно отметить, что с ростом количества сравниваемых выборок тех же объёмов мощность критерия относительно аналогичных конкурирующих гипотез, как правило, снижается, что абсолютно естественно. Например, сложнее выделить ситуацию и отдать предпочтение конкурирующей гипотезе, когда лишь одна из анализируемых выборок принадлежит некоторому другому закону.

Нельзя не отметить, что критерии Жанга со статистиками $Z_\text{К}$, $Z_\text{А}$, Z_C относительно некоторых альтернатив обладают заметным преимуществом в мощности.

6. Заключение

Построенные модели предельных распределений статистики (4) при использовании k-выборочного критерия однородности Андерсона–Дарлинга для анализа $k = 2 \div 11$ сравниваемых выборок (таблица 1) даёт возможность находить оценки p_{value}, что, несомненно, делает результаты статистических выводов более информативными и более обоснованными.

Такая же возможность реализована для предложенных критериев со статистиками вида (9) $S_{\max}^{Sm}$ (таблица 2), $S_{\max}^{LR}$ (таблица 3), $S_{\max}^{AD}$ (таблица 4).

Проведенный анализ мощности позволяет ориентироваться на наиболее предпочтительные критерии в зависимости от рассматриваемых альтернатив.

Литература

1. *Большев Л. Н.* Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1983. – 416 с.
2. *Lehmann E. L.* Consistency and unbiasedness of certain nonparametric tests / E. L. Lehmann // Ann. Math. Statist. 1951. Vol. 22, № 1. – P. 165–179.
3. *Rosenblatt M.* Limit theorems associated with variants of the von Mises statistic / M. Rosenblatt // Ann. Math. Statist. 1952. Vol. 23. – P. 617–623.
4. *Макаров А.А., Симонова Г.И.* Исследование мощности двухвыборочного критерия Андерсена–Дарлинга в случае засорения одной из выборок. // Статистические методы оценивания и проверки гипотез. Межвуз. сб. науч. тр. № 20, Перм. ун-т. 2007. – С. 40-52.
5. *Pettitt A.N.* A two-sample Anderson-Darling rank statistic // Biometrika. 1976. Vol. 63. No.1. P. 161-168.
6. *Scholz F.W., Stephens M.A.* K-Sample Anderson–Darling Tests // Journal of the American Statistical Association. 1987. Vol. 82. No. 399. – P. 918-924.
7. *Zhang J.* Powerful goodness-of-fit and multi-sample tests / J. Zhang // PhD Thesis. York University, Toronto. 2001. – 113 p. URL: http://www.collectionscanada.gc.ca/obj/s4/f2/dsk3/ftp05/NQ66371.pdf (дата обр. 28.01.2013).
8. *Zhang J.* Powerful Two-Sample Tests Based on the Likelihood Ratio / J. Zhang // Technometrics. 2006. V. 48. No. 1. – P.95-103.
9. *Zhang J., Wu Y.* k-Sample tests based on the likelihood ratio // Computational Statistics & Data Analysis. – 2007. – V. 51. – No. 9. – P. 4682-4691.

10. *Смирнов Н. В.* Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках / Н. В. Смирнов // Бюл. МГУ, Серия А. 1939. Т. 2, № 2. – С. 3–14.
11. *Kiefer J.* K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-v. Mises Tests // Annals of Mathematical Statistics, 1959. Vol. 30. No. 2. – P. 420-447.
12. *Лемешко Б.Ю.* О применении критериев проверки однородности законов распределения / Б.Ю. Лемешко, С.Б. Лемешко, И.В. Веретельникова // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2017. № 41. – С. 24-31.
13. *Lemeshko B. Y.* Application of Homogeneity Tests: Problems and Solution / B. Y. Lemeshko, I. V. Veretelnikova, S. B. Lemeshko, A. Y. Novikova // In: Rykov V., Singpurwalla N., Zubkov A. (eds) Analytical and Computational Methods in Probability Theory. ACMPT 2017. Lecture Notes in Computer Science. : monograph. - Cham : Springer, 2017. - 10684. - P. 461-475.
14. Лемешко Б.Ю. Критерии проверки гипотез об однородности. Руководство по применению. М: ИНФРА–М, 2017. – 207 с.
15. *Conover W. J.* Several k-sample Kolmogorov-Smirnov tests *// T*he Annals of Mathematical Statistics. 1965. Vol. 36, No. 3. – P.1019-1026.
16. *Conover W. J.* Practical Nonparametric Statistics / W. J. Conover. – 3d ed. – Wiley, 1999. – 584 p.
17. *Лемешко Б. Ю.* О сходимости распределений статистик и мощности критериев однородности Смирнова и Лемана–Розенблатта / Б. Ю. Лемешко, С. Б. Лемешко // Измерительная техника. 2005. № 12. – С. 9–14.

Лемешко Борис Юрьевич

Главный научный сотрудник кафедры прикладной и теоретической информатики НГТУ, д.т.н., профессор (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: Lemeshko@ami.nstu.ru, http://www.ami.nstu.ru/~headrd/

Веретельникова Ирина Викторовна

Аспирант кафедры прикладной и теоретической информатики НГТУ (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: ira-veterok@mail.ru.

About application and power of k-samples homogeneity tests

B. Yu. Lemeshko, I. V. Veretelnikova

Novosibirsk State Technical University

Properties of k-samples tests for homogeneity of distribution laws have been investigated. Tests have been proposed in which the statistics are being the maximum of the two-samples statistics of the Smirnov, the Lehmann-Rosenblatt and the Anderson-Darling tests of pairs of the k-samples. Models of limiting statistics distributions for the proposed tests, as well as for the Anderson-Darling k-sample test, have been made. Comparative analysis of the power of the tests has been done.

Keywords: k-samples tests, homogeneity test, power of test, statistical simulating