О k-выборочных критериях проверки однородности законов

Борис Юрьевич Лемешко, **Ирина Викторовна Веретельникова**

Новосибирский государственный технический университет

e-mail: Lemeshko@ami.nstu.ru

https://ami.nstu.ru/~headrd/

"Вычислительная и прикладная математика 2018", 08-10 октября 2018, ИВМ и МГ, Новосибирск, Россия

Введение

С необходимостью решения задач проверки гипотез о принадлежности двух (или более) выборок случайных величин одной и той же генеральной совокупности (проверки однородности) постоянно сталкиваются в различных приложениях. При этом возникают проблемы корректности применения и выбора наиболее предпочтительного критерия.

Задача проверки однородности k выборок формулируется следующим образом. Пусть x_{ij} j-е наблюдение 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 $H_0: F_1(x)=F_2(x)=\ldots=F_k(x)$ при любом x без указания общего для них закона распределения. Эмпирическую функцию распределения, соответствующую i-й выборке обозначим как $F_{in_i}(x)$.

На практике чаще всего применяются двухвыборочные критерии Смирнова [1] и Лемана–Розенблатта [1, 2, 3]. Значительно реже упоминается об использовании критерия Андерсона–Дарлинга [4] (Андерсона–Дарлинга–Петита) или его k-выборочного варианта [5], и ещё реже о применении k-выборочных вариантов критериев Смирнова или Лемана–Розенблатта [6, 7, 8]. Практически не говорится об использовании критериев однородности Жанга [9, 10].

Цель настоящей работы, являющейся развитием [11], состояла в исследовании распределений статистик и мощности критериев однородности при ограниченных объемах выборок, в уточнении объемов выборок, начиная с которых можно реально пользоваться предельными распределениями, в выяснении характера альтернатив, относительно которых тот или иной критерий имеет преимущество в мощности. При проведении исследований использовалась методика компьютерного моделирования и анализа статистических закономерностей, хорошо зарекомендовавшая себя в аналогичных работах [12, 13, 14, 15, 16, 17, 18], базирующаяся в основном на методе статистического моделирования.

1. Рассматриваемые критерии

1.1 Критерий Смирнова

Критерий однородности Смирнова предложен в работе [19]. 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

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

При практическом использовании критерия статистика D_{n_1,n_2} вычисляется в соответствии с соотношениями [1]:

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

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

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

При справедливости гипотезы H_0 статистика критерия Смирнова

$$S_C = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \tag{1}$$

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

Однако при ограниченных значениях n_1 и n_2 случайная величина D_{n_1,n_2} является дискретной, а число её возможных значений представляет собой наименьшее общее кратное n_1 и n_2 [1]. Ступенчатость условного распределения $G\left(S_C \middle| H_0\right)$ статистики S_C при равных n_1 и n_2 сохраняется даже при $n_i = 1000$. Поэтому предпочтительнее применять критерий, когда объемы выборок n_1 и n_2 не равны и представляют собой взаимно простые числа.

Другим недостатком критерия со статистикой (1) является то, что распределения $G\left(S_C \middle| H_0\right)$ с ростом n_1 и n_2 медленно приближаются к предельному распределению слева и при ограниченных n_1 и n_2 существенно отличаются от $K(s)$ (см. рис.1). В этой связи в [11] предложена простая модификация статистики (1):

$$S_{CM} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(D_{m,n} + \frac{n_1 + n_2}{4.6 n_1 n_2} \right),$$

у которой практически отсутствует последний недостаток.

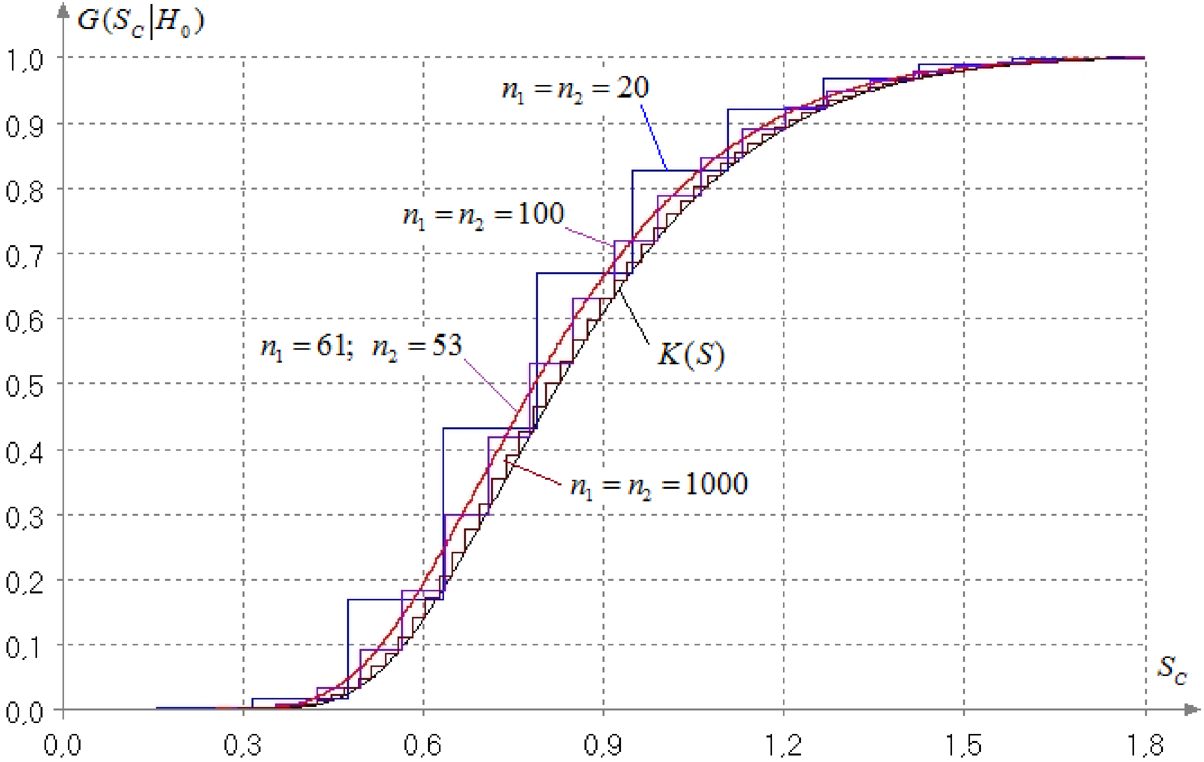

Рис. 1. Распределения статистики (1) при справедливости H_0 в зависимости от n_1 и n_2

1.2 Критерий Лемана–Розенблатта

Критерий однородности Лемана–Розенблатта представляет собой критерий типа ω^2. Критерий предложен в работе [2] и исследован в [3].

Статистика критерия используется в форме [1]

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

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 (4) в пределе распределена как $a1(t)$ [1].

В отличие от критерия Смирнова распределение статистики T быстро сходится к предельному $a1(T)$. При $n_1 = n_2 = 100$ распределение $G\left(T \middle| H_0\right)$ визуально совпадает с $a1(T)$, а при $n_1, n_2 \geq 45$ отклонением $G\left(T \middle| H_0\right)$ от $a1(T)$ на практике можно пренебречь.

1.3 Критерий Андерсона–Дарлинга

Двухвыборочный критерий Андерсона–Дарлинга (критерий однородности) рассмотрен в работе [4]. Статистика применяемого критерии определяется выражением

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

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 то же самое распределение $a2(t)$ [4], которое является предельным для статистики критерия согласия Андерсона–Дарлинга.

Сходимость распределения $G\left(A^2 \middle| H_0\right)$ статистики (3) к $a2(A^2)$ при ограниченных объёмах выборок была исследована в [20], где было показано, что при $n_1, n_2 \geq 45$ отклонение функции распределения $G\left(A^2 \middle| H_0\right)$ от $a2(A^2)$ не превышает 0.01.

1.4 Многовыборочный критерий Андерсона–Дарлинга

Многовыборочный вариант критерия согласия Андерсона–Дарлинга предложен в [5]. В предположении о непрерывности $F_i(x)$ по анализируемым выборкам строится объединённая общим объёмом $n=\sum_{i=1}^{k} n_i$ и упорядочивается $X_1 \le X_2 \le \ldots \le X_n$. Статистика критерия имеет вид [5]:

$$A_{kn}^2 = \frac{1}{n}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n-1}\frac{\left(nM_{ij} - jn_i\right)^2}{j(n-j)}, \tag{4}$$

где M_{ij} – число элементов в i-й выборке, которые не больше чем X_j. Проверяемая гипотеза H_0 отклоняется при больших значениях статистики (4).

В [5] таблица верхних процентных точек представлена не для статистики (4), а для статистики вида:

$$T_{kn} = \frac{A_{kn}^2 - (k-1)}{\sqrt{D[A_{kn}^2]}}. \tag{5}$$

Дисперсия статистики A_{kn}^2 определяется выражением [5]

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

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 (5) от числа сравниваемых выборок k иллюстрирует рис. 2. С ростом числа сравниваемых выборок это распределение медленно сходится к стандартному нормальному закону.

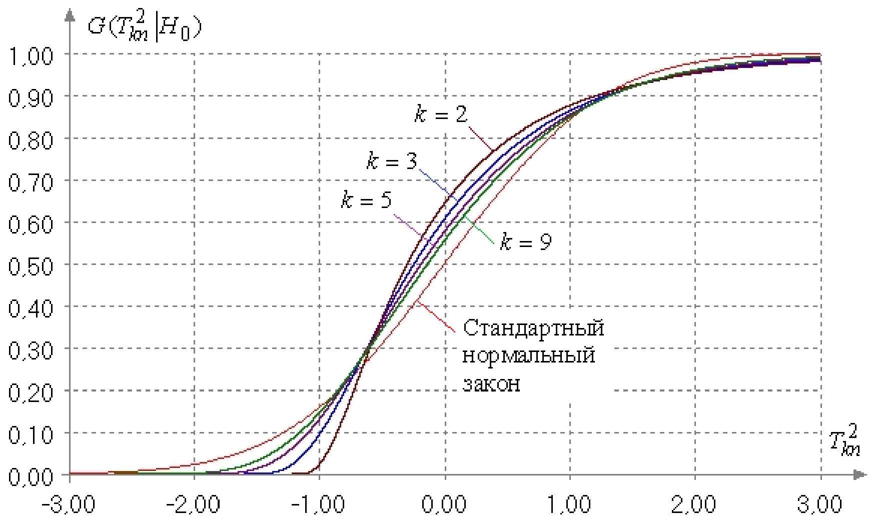

Рис. 2. Зависимость предельных распределений статистики (5) от числа сравниваемых выборок

Исследование распределений статистик методами статистического моделирования показало, что при использовании критериев отличие распределений статистик от соответствующих предельных не имеет практического значения при $n_i \geq 30$.

Таблица верхних процентных точек предельных распределений для статистики (5) представлена в [5]. Мы несколько уточнили и расширили таблицу критических значений.

Таблица 1

Уточненные верхние критические значения $T_{kn}^2(\alpha)$ статистики (5)

k	$1-\alpha$				
	0.75	0.90	0.95	0.975	0.99
2	0.325	1.228	1.966	2.731	3.784
3	0.439	1.300	1.944	2.592	3.429
4	0.491	1.321	1.925	2.511	3.277
5	0.523	1.331	1.900	2.453	3.153
6	0.543	1.333	1.885	2.410	3.078
7	0.557	1.337	1.870	2.372	3.017
8	0.567	1.335	1.853	2.344	2.970
9	0.577	1.334	1.847	2.323	2.927
10	0.582	1.3345	1.838	2.306	2.899
11	0.589	1.332	1.827	2.290	2.867
∞	0.674	1.282	1.645	1.960	2.326

Одновременно для предельных распределений статистики (5) были построены приближенные модели законов (для $k = 2 \div 11$).

Хорошими моделями оказались [21] законы семейства бета-распределений III рода с плотностью

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

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

Представленные в таблице 2 модели $\mathrm{B_{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ с приведенными значениями параметров позволяют по значениям статистики, вычисленным по соотношению (5), находить оценки p_{value} при соответствующем числе k сравниваемых выборок.

Т а б л и ц а 2

Модели предельных распределений статистики (11)

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)

1.5 Критерии однородности Жанга

Предложенные Жангом критерии однородности [9, 10] являются развитием критериев Смирнова, Лемана–Розенблатта и Андерсона–Дарлинга и дают возможность сравнивать $k \geq 2$ выборок.

Критерии согласия Жанга [9] показывают некоторое преимущество в мощности по сравнению с критериями согласия Крамера–Мизеса–Смирнова и Андерсона–Дарлинга [22], но недостатком, ограничивающим применение критериев согласия Жанга, является зависимость распределений статистик от объёмов выборок.

Этим же недостатком обладают варианты критериев Жанга для проверки однородности законов. Для преодоления этого недостатка автор [9] предлагает для оценивания p_{value} использовать метод Монте–Карло.

Задача моделирования распределений статистик критериев однородности Жанга, по сравнению с аналогичной задачей для критериев согласия, оказывается много проще, так как приходится моделировать распределения статистик $G(S|H_0)$ критериев в случае принадлежности анализируемых выборок равномерному закону.

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 $\hat{F}(X_m) = (m - 0.5) / n$ [9].

Статистика Z_K критерия однородности Жанга имеет вид [9]:

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

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). Распределения статистики зависят от n_i и от k. На принятие решения влияет дискретность статистики, которая с ростом k становится менее выраженной (см. рис.3).

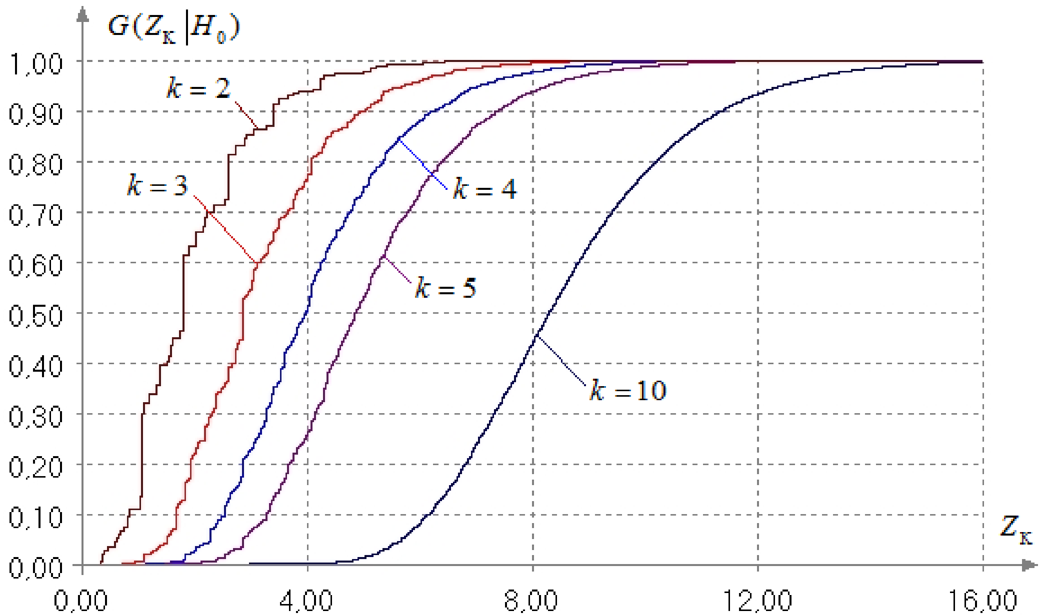

Рис. 3. Зависимость распределений статистики (6) от k при $n_i = 20$

Статистика Z_{A} критерия однородности Жанга определяется выражением [9]:

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

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

Критерий левосторонний: проверяемая гипотеза H_0 отклоняется при **малых** значениях статистики (7).

Распределения статистики зависят от n_i и от k (рис. 4 и 5).

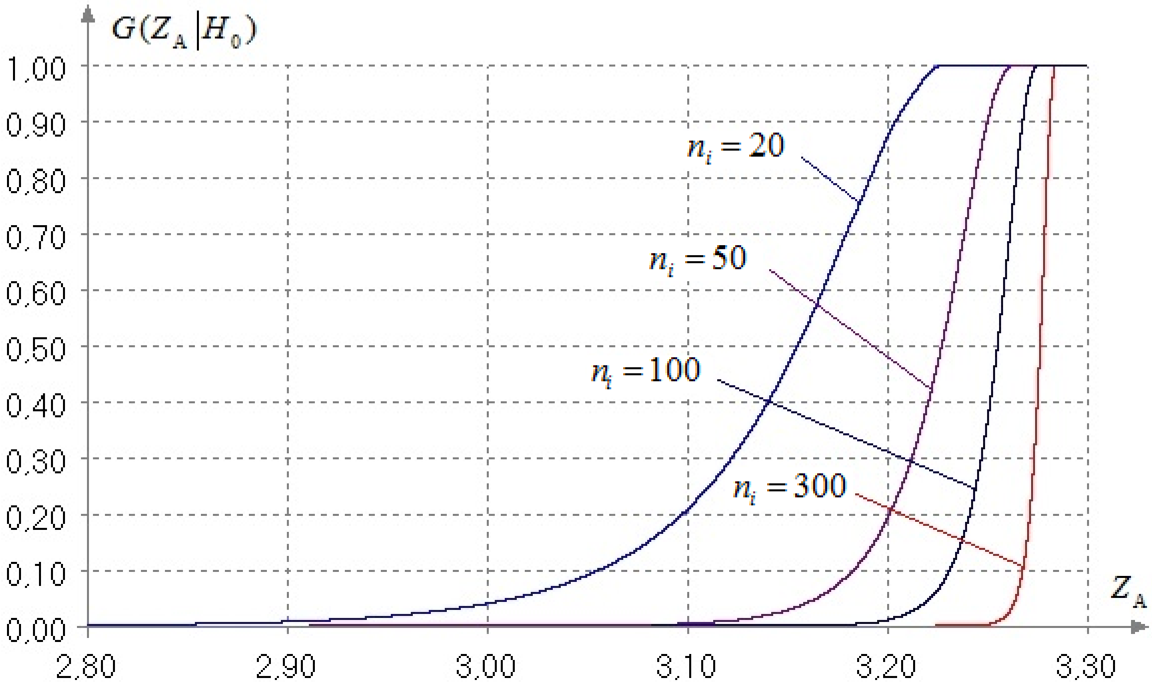

Рис. 4. Зависимость распределений статистики (7) от объёмов выборок $k = 2$ при равных n_i

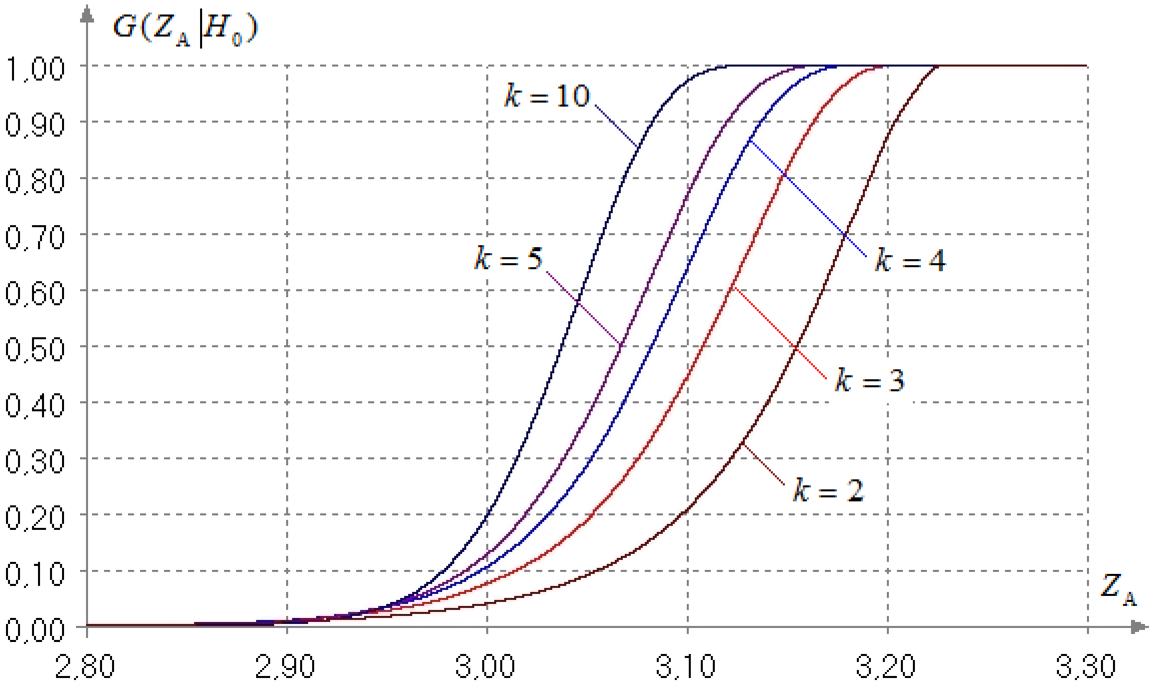

Рис. 5. Зависимость распределений статистики (7) от k при $n_i = 20$

Статистика Z_C критерия однородности Жанга k выборок вычисляется в соответствии с выражением [9]:

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

Критерий также левосторонний: проверяемая гипотеза H_0 отклоняется при **малых** значениях статистики (8). Распределения статистики зависят от n_i и от k, зависимость от k показана на рис. 6.

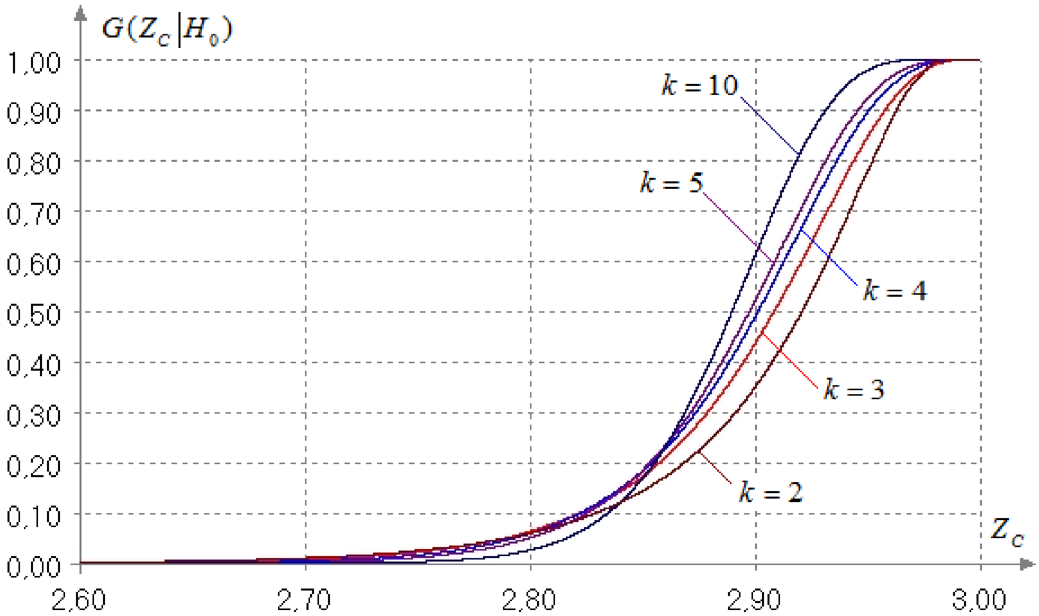

Рис. 6. Зависимость распределений статистики (8) от k при $n_i = 20$

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.

2. Двухвыборочные критерии при анализе k выборок

Различные подходы к построению k -выборочных аналогов критериев Смирнова, Лемана–Розенблатта и Андерсона–Дарлинга рассматривались в работе [6].

k -выборочный вариант критерия Колмогорова–Смирнова, основанный на таком подходе, был построен в [7] и рассматривается в последовательных изданиях книги [8].

На использовании такого же подхода построен k -выборочный критерий Андерсона–Дарлинга [5], рассмотренный выше, для которого нами были построены модели предельных распределений.

Во всех 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

$$S_{\max} = \max_{\substack{1 \le i \le k \\ i < j \le k}} \{S_{i,j}\}, \tag{2.17}$$

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), 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.

2.1. k–выборочный критерий Смирнова (max)

В качестве $S_{i,j}$ в (2.17) в этом случае будет рассматриваться модификация статистики Смирнова (2.3), 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. 2.28) и отличаются от асимптотических (предельных) распределений (см. рис. 2.29). 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.

Верхние критические значения для статистики $S_{\max}^{Sm}$, построенные по эмпирическим распределениям статистик, полученным методом Монте–Карло при количестве имитационных экспериментов $N=10^6$, представлены в таблице 2.20, а в таблице 2.21 приведены построенные модели асимптотических (предельных) распределений статистики $S_{\max}^{Sm}$ при числе сравниваемых выборок $k=3\div 11$.

Представленные в таблице 2.21 модели $\mathrm{B_{III}}(\theta_0,\theta_1,\theta_2,\theta_3,\theta_4)$ бета-распределения 3-го рода (2.13) с приведенными значениями параметров, позволяют по значениям статистики, вычисленным в соответствии с соотношением (2.17) с использованием в качестве $S_{i,j}$ статистики Смирнова (2.1) или её модификации (2.3), находить оценки p_{value} при соответствующем числе k сравниваемых выборок.

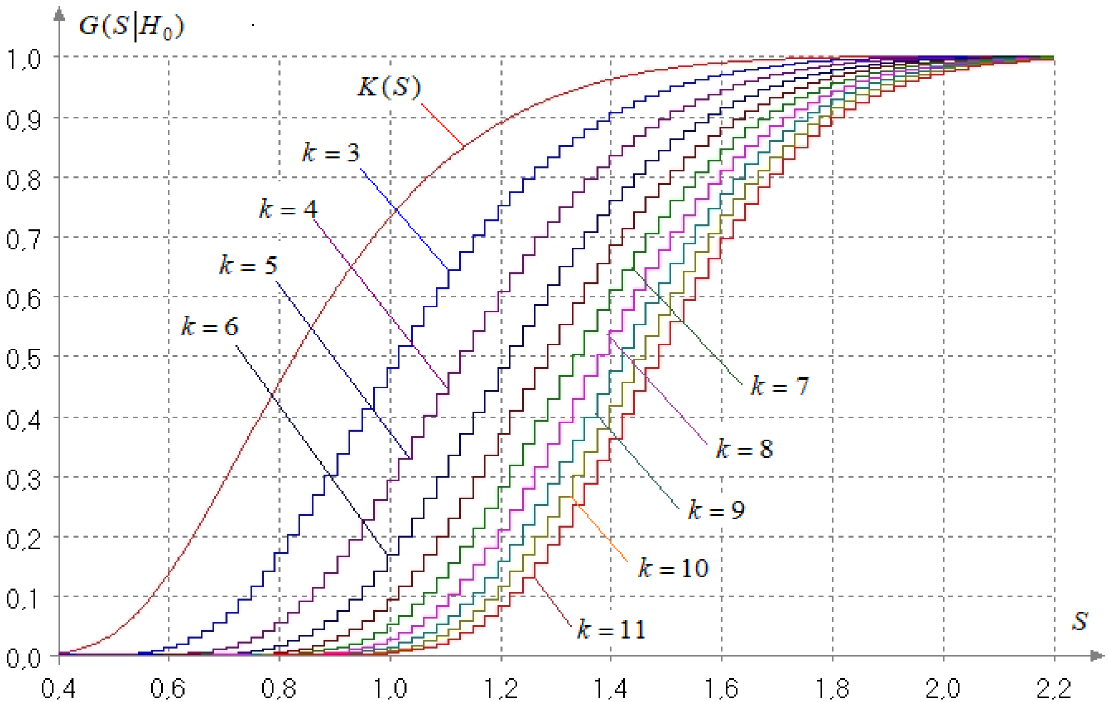

Рис. 2.28. Распределения статистики $S_{\max}^{Sm}$ k-выборочного критерия Смирнова при $n_i = 1000$, $i = \overline{1,k}$

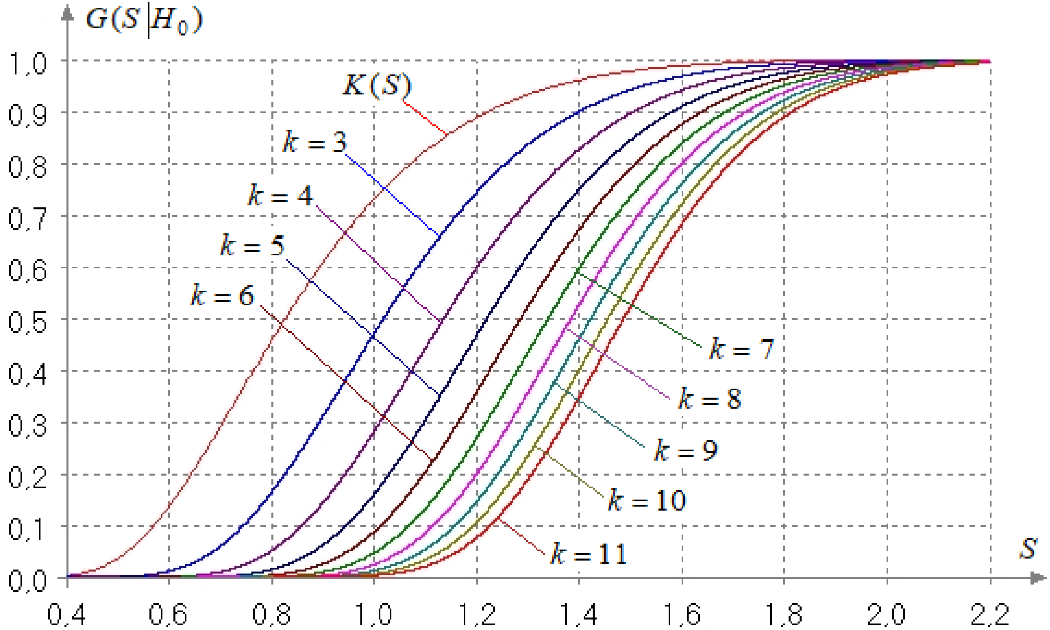

Рис. 2.29. Асимптотические распределения статистики $S_{\max}^{Sm}$ k -выборочного критерия Смирнова

Т а б л и ц а 2.20

Верхние критические значения статистики $S_{\max}^{Sm}$ Смирнова

k	$1-\alpha$				
	0.75	0.90	0.95	0.975	0.99
2	1.0192	1.2238	1.3581	1.4802	1.6276
3	1.2059	1.4006	1.5266	1.6396	1.7782
4	1.3194	1.5070	1.6278	1.7372	1.8701
5	1.4002	1.5822	1.6997	1.8057	1.9329
6	1.4620	1.6396	1.7545	1.8573	1.9833
7	1.5127	1.6867	1.7989	1.9009	2.0245
8	1.5546	1.7264	1.8362	1.9362	2.0596
9	1.5911	1.7598	1.8682	1.9672	2.0879
10	1.6224	1.7893	1.8969	1.9940	2.1148
11	1.6506	1.8156	1.9217	2.0182	2.1371

Т а б л и ц а 2.21

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)

2.6.2. k–выборочный критерий Лемана–Розенблатта (max)

В качестве $S_{i,j}$ в статистике $S_{\max}^{LR}$ вида (2.17) в этом случае используется статистика (2.4) Лемана–Розенблатта. Зависимость распределений статистики при справедливости H_0 от числа выборок иллюстрирует рис. 2.30.

Верхние критические значения для статистики $S_{\max}^{LR}$, построенные по результатам статистического моделирования при количестве имитационных экспериментов $N = 10^6$, представлены в таблице 2.24.

Построенные модели асимптотических (предельных) распределений статистики $S_{\max}^{LR}$ при числе сравниваемых выборок $k = 3 \div 11$ представлены в таблице 2.25.

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 2.25 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}.

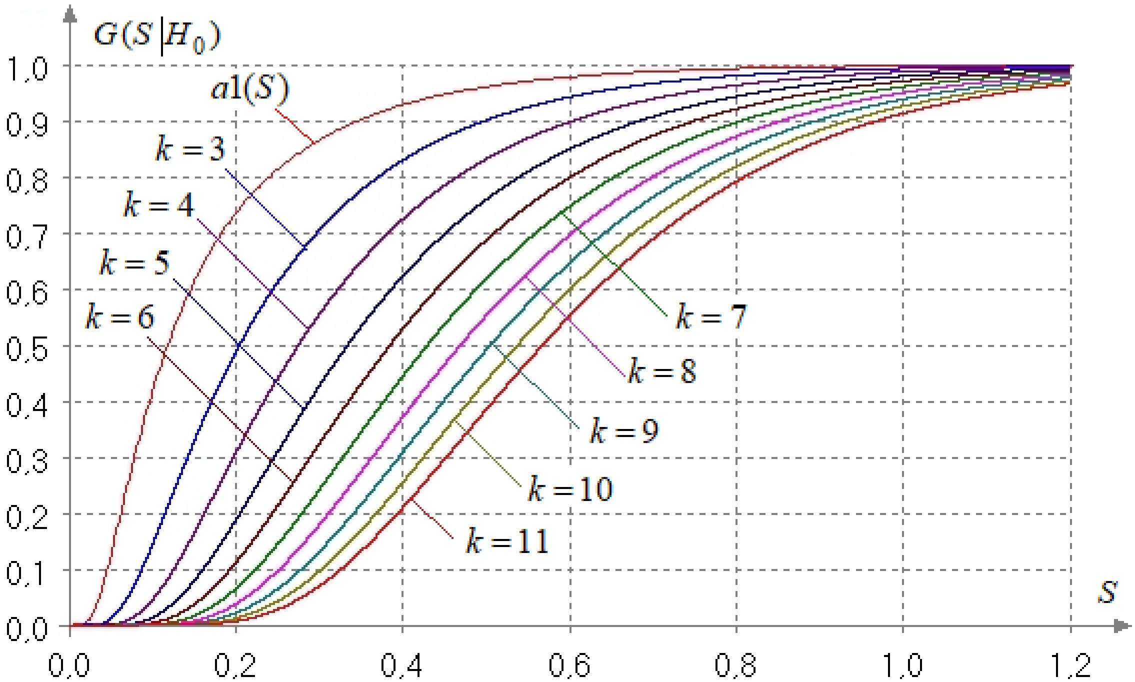

Рис. 2.30. Распределения статистики $S_{\max}^{LR}$ k-выборочного критерия Лемана–Розенблатта

Таблица 2.24

Верхние критические значения статистики $S_{\max}^{LR}$ Лемана–Розенблатта

k	$1-\alpha$				
	0.75	0.90	0.95	0.975	0.99
2	0.2094	0.3473	0.4614	0.5806	0.7435
3	0.3306	0.4995	0.6283	0.7581	0.9308
4	0.4206	0.6050	0.7413	0.8770	1.0550
5	0.4924	0.6856	0.8267	0.9653	1.1429
6	0.5521	0.7512	0.8959	1.0365	1.2175
7	0.6037	0.8072	0.9524	1.0947	1.2774
8	0.6481	0.8550	1.0015	1.1444	1.3298
9	0.6876	0.8976	1.0457	1.1902	1.3781
10	0.7237	0.9355	1.0858	1.2303	1.4171
11	0.7563	0.9703	1.1214	1.2655	1.4535

Таблица 2.25

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)

2.6.3. k–выборочный критерий Андерсона–Дарлинга (max)

В статистике $S_{\max}^{AD}$ вида (2.17) качестве $S_{i,j}$ используется статистика (2.7) Андерсона–Дарлинга. Зависимость распределений статистики $S_{\max}^{AD}$ при справедливости H_0 от числа выборок иллюстрирует рис. 2.31.

Верхние критические значения для статистики $S_{\max}^{AD}$, построенные по результатам статистического моделирования при количестве имитационных экспериментов $N = 10^6$, представлены в таблице 2.28.

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 2.29.

В этом случае лучшими моделями оказались бета-распределения 3-го рода (2.13), которые в виде $\mathrm{B_{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ с конкретными значениями параметров приведены в таблице 2.29 и могут использоваться для оценки p_{value} при k сравниваемых выборках.

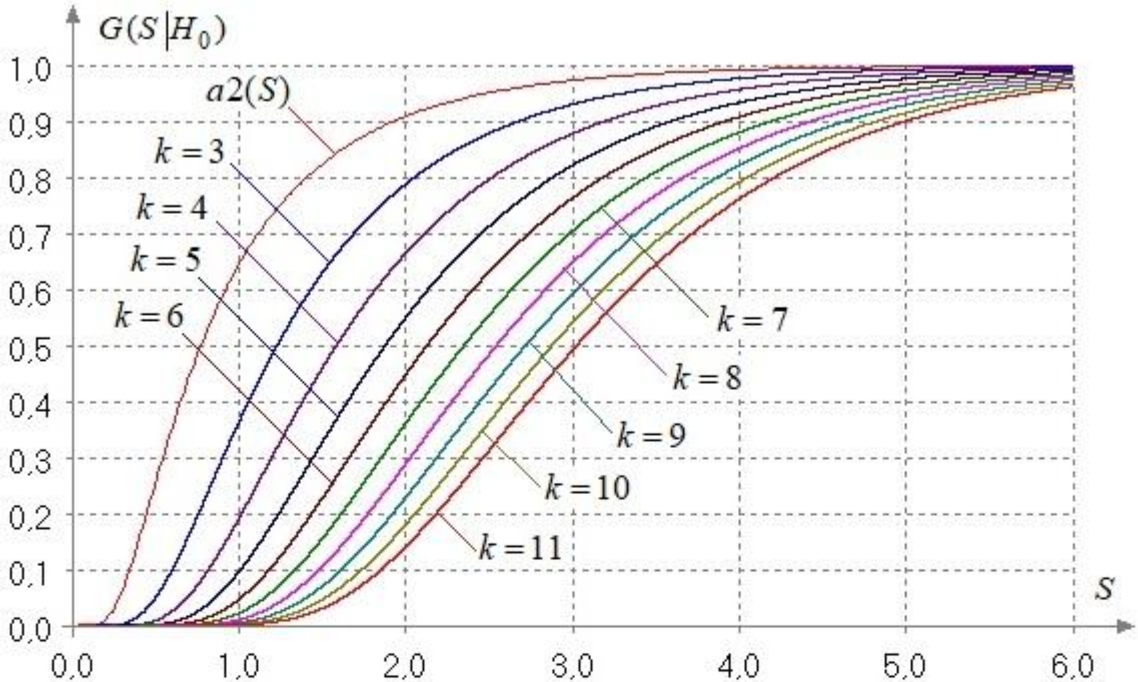

Рис. 2.31. Распределения статистики $S_{\max}^{AD}$ k -выборочного критерия Андерсона–Дарлинга

Таблица 2.28

Верхние критические значения статистики $S_{\max}^{AD}$ Андерсона–Дарлинга

k	$1-\alpha$				
	0.75	0.90	0.95	0.975	0.99
2	1.2479	1.9330	2.4924	3.0775	3.8781
3	1.8535	2.6796	3.31215	3.95176	4.7924
4	2.2990	3.1966	3.8682	4.5368	5.4076
5	2.6514	3.5948	4.2877	4.9686	5.8472
6	2.9431	3.9187	4.6292	5.3175	6.2089
7	3.1943	4.1950	4.9097	5.6063	6.5118
8	3.4135	4.4292	5.1501	5.8531	6.7710
9	3.6094	4.6395	5.3660	6.0733	7.0076
10	3.7867	4.8270	5.5616	6.2720	7.2042
11	3.9466	4.9974	5.7384	6.4512	7.3837

Таблица 2.29

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)

2.7. Критерий однородности χ^2

Пусть имеется k выборок $x_{i1}, x_{i2}, \ldots, x_{in_i}$, $i = \overline{1,k}$, непрерывных случайных величин и $n = \sum_{i=1}^{k} n_i$. Общая область, которой принадлежат выборки, разбивается на r интервалов (групп). Пусть η_{ij} – количество элементов i-й выборки, попавших в j-й интервал, тогда $n_i = \sum_{j=1}^{r} \eta_{ij}$.

Статистика критерия однородности χ^2 имеет вид

$$\chi^2 = n \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(\eta_{ij} - \nu_j n_i / n)^2}{\nu_j n_i} = n \left(\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\eta_{ij}^2}{\nu_j n_i} - 1 \right), \quad (2.19)$$

где $\nu_j = \sum_{l=1}^{k} \eta_{lj}$ – общее число элементов всех выборок, попавших j-й интервал.

Асимптотическим распределением статистики (2.19) является χ^2-распределение с числом степеней свободы $(k-1)(r-1)$.

2.7. Примеры применения

Рассмотрим применение рассмотренных в разделе критериев проверки однородности законов на примере анализа 3-х нижеприведенных выборок, каждой объёмом в 40 наблюдений.

0.321	0.359	-0.341	1.016	0.207	1.115	1.163	0.900	-0.629	-0.524
-0.528	-0.177	1.213	-0.158	-2.002	0.632	-1.211	0.834	-0.591	-1.975
-2.680	-1.042	-0.872	0.118	-1.282	0.766	0.582	0.323	0.291	1.387
-0.481	-1.366	0.351	0.292	0.550	0.207	0.389	1.259	-0.461	-0.283

0.890	-0.700	0.825	1.212	1.046	0.260	0.473	0.481	0.417	1.825
1.841	2.154	-0.101	1.093	-1.099	0.334	1.089	0.876	2.304	1.126
-1.134	2.405	0.755	-1.014	2.459	1.135	0.626	1.283	0.645	1.100
2.212	0.135	0.173	-0.243	-1.203	-0.017	0.259	0.702	1.531	0.289

0.390	0.346	1.108	0.352	0.837	1.748	-1.264	-0.952	0.455	-0.072
-0.054	-0.157	0.517	1.928	-1.158	-1.063	-0.540	-0.076	0.310	-0.237
-1.109	0.732	2.395	0.310	0.936	0.407	-0.327	1.264	-0.025	-0.007
0.164	0.396	-1.130	1.197	-0.221	-1.586	-0.933	-0.676	-0.443	-0.101

Эмпирические распределения, соответствующие данным выборкам, представлены на рис. 2.32

В данном случае 1-я и 3-я выборки были смоделированы в соответствии со стандартным нормальным законом, а полученные значения псевдослучайных величин округлены до 3-х значащих цифр после десятичной точки. В то время как 2-я выборка получена в соответствии с нормальным законом с параметром сдвига 0.5 и стандартным отклонением 1.1.

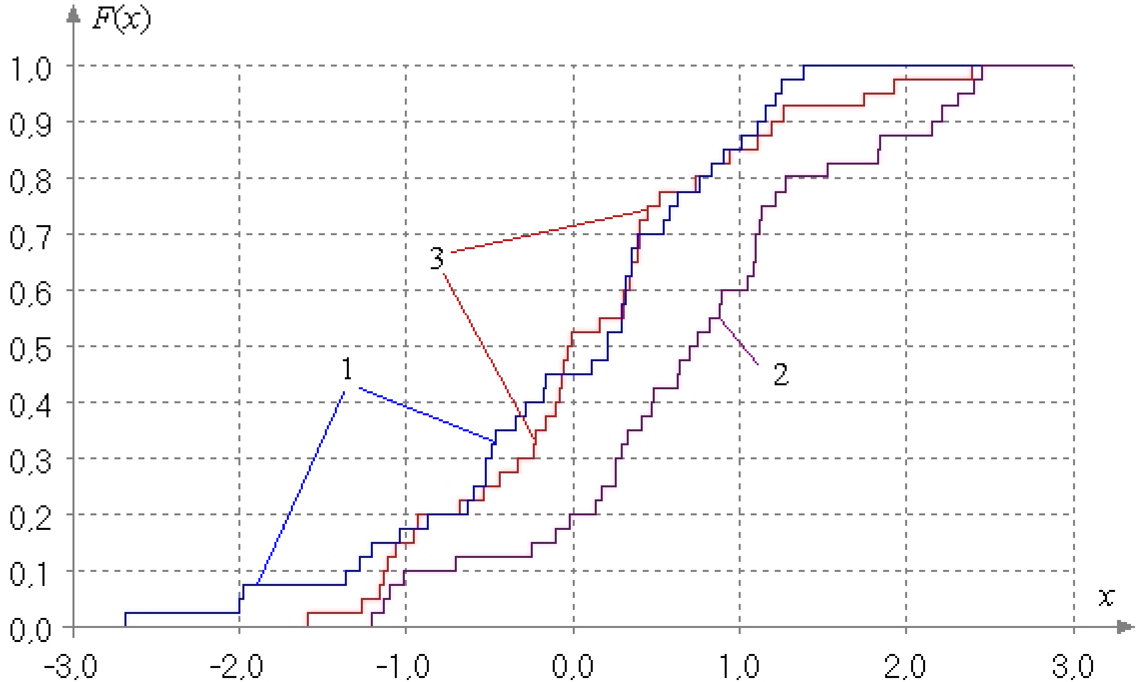

Рис. 2.32. Эмпирические распределения, соответствующие выборкам

Проверим гипотезу об однородности 1-й и 2-й выборок. В таблице 2.35 приведены результаты проверки: значения статистик критериев и достигнутые уровни значимости p_{value}. Оценки p_{value} вычислялись по значению статистики в соответствии с распределением (2.8) для критерия Андерсона–Дарлинга, в соответствии с распределением (2.5) для критерия Лемана–Розенблатта, в соответствии с распределением (2.2) для критерия Смирнова, в соответствии с бета-распределением 3-го рода из таблицы 2.8 при $k=2$ для k-выборочного критерия Андерсона–Дарлинга. Распределения статистик (2.12), (2.13) и (2.14) критериев Жанга и оценки p_{value} находились в результате моделирования. По приведенным оценкам p_{value} очевидно, что гипотеза об однородности по всем критериям должна быть отклонена.

Таблица 2.35

Результаты проверки однородности 1-й и 2-й выборок

Критерий	Статистика	p_{value}
Андерсона–Дарлинга	5.19801	0.002314
k-выборочный Андерсона–Дарлинга	5.66112	0.003260
Лемана–Розенблатта	0.9650	0.002973
Смирнова	1.56525	0.014893
Смирнова модифицированный	1.61386	0.010933
Жанга Z_{A}	2.99412	0.0007
Жанга Z_C	2.87333	0.0008
Жанга Z_{K}	5.58723	0.0150
χ^2, $r=10$	16.3254	0.06039
χ^2, $r=8$	13.3293	0.06448

В таблице 2.36 приведены результаты проверки гипотезы об однородности 1-й и 3-й выборок. Здесь оценки p_{value} по всем критериям очень высокие, поэтому проверяемая гипотеза об однородности не должна отклоняться.

Таблица 2.36

Результаты проверки однородности 1-й и 3-й выборок

Критерий	Статистика	p_{value}
Андерсона–Дарлинга	0.49354	0.753415
k-выборочный Андерсона–Дарлинга	-0.68252	0.767770
Лемана–Розенблатта	0.0500	0.876281
Смирнова	0.447214	0.989261
Смирнова модифицированный	0.495824	0.966553
Жанга Z_{A}	3.1998	0.332
Жанга Z_C	3.07077	0.384
Жанга Z_{K}	1.7732	0.531
χ^2, $r=10$	11.6508	0.23372
χ^2, $r=8$	4.0000	0.7798

В таблице 2.37 показаны результаты проверки гипотезы об однородности трёх рассматриваемых выборок по k-выборочному критерию Андерсона–Дарлинга, по критериям Жанга и по критериям со статистиками $S_{\max}^{Sm}$, $S_{\max}^{LR}$, $S_{\max}^{AD}$.

Таблица 2.37

Результаты проверки однородности 3-х выборок

Критерий	Статистика	p_{value}
k-выборочный Андерсона–Дарлинга	4.73219	0.0028
Жанга Z_A	3.02845	0.0016
Жанга Z_C	2.92222	0.0017
Жанга Z_K	7.00231	0.0218
max Андерсона–Дарлинга	5.19801	0.0064
max Лемана–Розенблатта	0.9650	0.0094
max Смирнова модифицированный	1.72566	0.0144
χ^2, $r=10$	25.556	0.1104
χ^2, $r=8$	19.200	0.1574

Заметим, что критерии Андерсона–Дарлинга $S_{\max}^{AD}$ и Лемана–Розенблатта $S_{\max}^{LR}$ зафиксировали максимальное отклонение между 1-й и 2-й выборками, а критерий Смирнова $S_{\max}^{Sm}$ – между 2-й и 3-й. Общий результат показывает, что проверяемая гипотеза об однородности 3-х выборок должна быть отклонена.

Можно обратить внимание на существенную зависимость результатов проверки по критерию однородности χ^2 от выбираемого числа интервалов r.

2.8. Сравнительный анализ мощности критериев

В случае k выборок относительно изменения параметра сдвига порядок предпочтения критериев имеет вид [23]:

$$S_{\max}^{AD} \succ \text{Андерсона–Дарлинга} \succ S_{\max}^{LR} \succ S_{\max}^{Sm} \succ \text{Жанга } Z_C \succ$$
$$\text{Жанга } Z_{\text{A}} \succ \text{Жанга } Z_{\text{K}} \succ \chi^2 .$$

Относительно изменения параметра масштаба –

$$\text{Жанга } Z_C \succ \text{Жанга } Z_{\text{A}} \succ \text{Жанга } Z_{\text{K}} \succ \text{Андерсона–Дарлинга} \succ$$
$$\chi^2 \succ S_{\max}^{AD} \succ S_{\max}^{Sm} \succ S_{\max}^{LR} .$$

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

Относительно ситуации, когда все выборки (кроме одной) принадлежат нормальному закону, а одна – логистическому, критерии располагаются по мощности в следующем порядке:

$$\text{Жанга } Z_{\text{A}} \succ \text{Жанга } Z_C \succ \text{Жанга } Z_{\text{K}} \succ \chi^2 \succ \text{Андерсона–Дарлинга} \succ S_{\max}^{Sm} \succ S_{\max}^{AD} \succ S_{\max}^{LR} .$$

References

1. Bol’shev, L. N., Smirnov, N. V. Tables for Mathematical Statistics [in Russian], Nauka, Moscow. 1983. – 416 p.
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. *Pettitt A.N.* A two-sample Anderson-Darling rank statistic // Biometrika. 1976. Vol. 63. No.1. P. 161-168.
5. *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.
6. *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.
7. *Conover W. J.* Several k-sample Kolmogorov-Smirnov tests // The Annals of Mathematical Statistics. 1965. Vol. 36, No. 3. – P.1019-1026.
8. Conover, W. J. Practical Nonparametric Statistics, 3rd ed. New York: John Wiley and Sons. 1999. 592 p.
9. *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 обращения 28.01.2013).
10. *Zhang J.* Powerful Two-Sample Tests Based on the Likelihood Ratio / J. Zhang // Technometrics. – 2006. – V. 48. – No. 1. – P.95-103. DOI 10.1198/004017005000000328
11. *Lemeshko B. Yu.* Statistical distribution convergence and homogeneity test power for Smirnov and Lehmann–Rosenblatt tests / B. Yu. Lemeshko, S. B. Lemeshko // Measurement Techniques. – 2005. – Vol. 48, № 12. – P. 1159–1166.
12. Statistical Data Analysis, Simulation and Study of Probability Regularities. Computer Approach: monograph / B.Yu. Lemeshko, S.B. Lemeshko, S.N. Postovalov, E.V. Chimitova. – Novosibirsk. NSTU Publisher, 2011. – 888 p.
13. *Lemeshko B. Yu.* Power and robustness of criteria used to verify the homogeneity of means / B. Yu. Lemeshko, S. B. Lemeshko // Measurement Techniques. – 2008. – Vol. 51, № 9. – P. 950–959.
14. *Lemeshko B.* Bartlett and Cochran tests in measurements with probability laws different from normal / B. Lemeshko, E. Mirkin // Measurement Techniques. – 2004. – Vol. 47, № 10. – P. 960–968.

15. *Lemeshko B.Yu.* Application and power of criteria for testing the homogeneity of variances. Part I. Parametric criteria / B. Yu. Lemeshko, S. B. Lemeshko, A. A. Gorbunova // Measurement Techniques. – 2010. – Vol. 53, № 3. – P. 237–246.

16. *Lemeshko B.Yu., Lemeshko S.B., and A.A. Gorbunova*. Application and power of criteria for testing the homogeneity of variances. Part II. Nonparametric criteria // Measurement Techniques, Vol. 53, No. 5, 2010. P.476-486.

17. *Lemeshko B.Y., Sataeva T.S.* Application and Power of Parametric Criteria for Testing the Homogeneity of Variances. Part III //Measurement Techniques, 2017. Vol. 60. No. 1. – P. 7-14.

18. *Lemeshko B.Y., Sataeva T.S.* Application and Power of Parametric Criteria for Testing the Homogeneity of Variances. Part IV // Измерительная техника. 2017. № 5. – P. 12-17.

19. *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.

20. Postovalov S.N. Using of computer modeling for expanding application of classical methods of statistics hypothesis checking. DEA Thesis. Novosibirsk. 2013. 298 p.

21. *Lemeshko B.Yu* Tests for homogeneity. Guide on the application. M: INFRA–M, 2017. – 208 p. DOI: 10.12737/22368

22. *Lemeshko B.Yu* Nonparametric goodness-of-fit tests. Guide on the application. M: INFRA–M, 2014. –163 p. DOI: 10.12737/11873

23. *Лемешко Б. Ю.* Мощность *k*-выборочных критериев проверки однородности законов / Б. Ю. Лемешко, И. В. Веретельникова // Измерительная техника. 2018. № 7. – С. 3-7.

Исследования выполнены 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).

Спасибо за внимание!

Текущая версия программной системы ISW.

Доступна по адресу: https://ami.nstu.ru/~headrd/