Математические методы исследования

УДК 519.234

СОСТОЯТЕЛЬНЫЕ КРИТЕРИИ ПРОВЕРКИ АБСОЛЮТНОЙ ОДНОРОДНОСТИ НЕЗАВИСИМЫХ ВЫБОРОК

© А. И. Орлов[1]

Статья поступила 18 октября 2010 г.

Рассмотрены критерии проверки гипотезы абсолютной однородности двух независимых выборок. Установлено, что состоятельными являются только критерии Смирнова и Лемана – Розенблатта. Приведены алгоритмы расчетов и правила принятия решений для этих критериев.

Ключевые слова: непараметрические статистические критерии; проверка однородности; критерий Смирнова; критерий типа омега-квадрат.

Как и в предыдущих работах [1 – 3], примем следующую модель порождения данных. Элементы первой выборки $x_1, x_2, ..., x_m$ будем рассматривать как результаты m независимых наблюдений некоторой числовой случайной величины X с функцией распределения $F(x)$, неизвестной статистику, а элементы второй выборки $y_1, y_2, ..., y_n$ — как результаты n независимых наблюдений другой случайной величины Y с функцией распределения $G(x)$, также неизвестной статистику. Предположим также, что наблюдения в одной выборке не зависят от наблюдений в другой, поэтому выборки и называются независимыми.

Понятие «однородность», т.е. «отсутствие различия», может быть формализовано в терминах вероятностной модели различными способами. Наивысшая степень однородности (абсолютная однородность) достигается, если обе выборки взяты из одной и той же генеральной совокупности, т.е. справедлива нулевая гипотеза

$$H_0:\ F(x) = G(x) \text{ при всех } x.$$

Отсутствие абсолютной однородности означает, что верна альтернативная гипотеза, согласно которой

$$H_1: F(x_0) \neq G(x_0)$$

хотя бы при одном значении аргумента x_0. Если гипотеза H_0 принята, то выборки можно объединить в одну, если нет — то нельзя.

В некоторых случаях целесообразно проверять совпадение не функций распределения, а лишь некоторых характеристик случайных величин X и Y — математических ожиданий, медиан, дисперсий, коэффициентов вариации и др. [2, 3].

В соответствии с методологией прикладной статистики естественно потребовать, чтобы рекомендуемый для массового использования статистический критерий абсолютной однородности был состоятельным, т.е. для любых отличных друг от друга функций распределения $F(x)$ и $G(x)$ (другими словами, при справедливости альтернативной гипотезы H_1) вероятность отклонения гипотезы H_0 стремилась к единице при увеличении объемов выборок m и n. Из перечисленных в работах [1, 3] критериев однородности состоятельными являются только двухвыборочные критерии Смирнова и омега-квадрат (Лемана – Розенблатта). Проведенное в Институте высоких статистических технологий и эконометрики МГТУ им. Н. Э. Баумана исследование мощности (методом статистических испытаний) критериев однородности (при различных вариантах функций распределения $F(x)$ и $G(x)$) подтвердило преимущество критериев Смирнова и омега-квадрат и при малых объемах выборок (6 – 12). Однако указанный вывод не был подкреплен алгоритмами расчетов и правилами принятия решений для отмеченных критериев. Этим вопросам и посвящена данная работа.

Критерий Смирнова однородности двух независимых выборок

Критерий предложен чл.-корр. АН СССР Н. В. Смирновым в 1939 г. [4]. Единственное ограничение — функции распределения $F(x)$ и $G(x)$ должны быть непрерывными.

Согласно Л. Н. Большеву и Н. В. Смирнову [4], значение эмпирической функции распределения в точке x равно доле результатов наблюдений в выборке, меньших x. Критерий Смирнова основан на использовании эмпирических функций распределения $F_m(x)$ и $G_n(x)$, построенных по первой и второй выборкам со-

[1] Институт высоких статистических технологий и эконометрики Московского государственного технического университета им. Н. Э. Баумана, Москва, Россия;
e-mail: prof-orlov@mail.ru

Таблица 1. Данные для расчета значения статистики $D^{+}_{m,n}$

Номер элемента в объединенной выборке	Элементы выборок x	Номера выборок	$F_m(x)$	r/n	$r/n - F_m(x)$	$G_n(x)$	$\frac{s-1}{m}$	$G_n(x) - \frac{s-1}{m}$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	1	0			0	0	0
2	1	2	0,083	0,071	–0,012	0		
3	2	1	0,083			0,071	0,083	–0,012
4	2	2	0,083	0,143	0,06	0,071		
5	3	1	0,167			0,143	0,167	–0,024
6	5	1	0,25			0,143	0,25	–0,107
7	6	2	0,333	0,214	–0,119	0,143		
8	7	1	0,333			0,214	0,333	–0,119
9	7	2	0,333	0,286	–0,047	0,214		
10	11	2	0,417	0,357	–0,06	0,286		
11	13	1	0,417			0,357	0,417	–0,06
12	14	1	0,5			0,357	0,5	–0,143
13	15	1	0,583			0,357	0,583	–0,226
14	15	2	0,583	0,429	–0,154	0,357		
15	15	2	0,583	0,5	–0,083	0,357		
16	17	1	0,667			0,5	0,667	–0,167
17	21	2	0,75	0,571	–0,179	0,5		
18	22	1	0,75			0,571	0,75	–0,179
19	25	2	0,833	0,643	–0,19	0,571		
20	29	2	0,833	0,714	–0,119	0,643		
21	30	2	0,833	0,786	–0,047	0,714		
22	33	2	0,833	0,857	0,024	0,786		
23	44	2	0,833	0,929	0,096	0,857		
24	47	2	0,833	1,0	0,167	0,929		
25	66	1	0,833			1,0	0,833	0,167
26	97	1	0,917			1,0	0,917	0,083

ответственно. Значение двухвыборочной статистики Смирнова

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|$$

сравнивают с соответствующим критическим значением [4] и по результатам сравнения принимают или отклоняют гипотезу H_0 о совпадении (однородности) функций распределения. Практически значение двухвыборочной статистики Смирнова $D_{m,n}$ рекомендуется [4] вычислять по формулам

$$D^{+}_{m,n} = \max_{1 \le r \le n} \left[\frac{r}{n} - F_m(y'_r) \right] = \max_{1 \le r \le m} \left[G_n(x'_s) - \frac{s-1}{m} \right], \quad (1)$$

$$D^{-}_{m,n} = \max_{1 \le r \le n} \left[F_m(y'_r) - \frac{r-1}{n} \right] = \max_{1 \le r \le m} \left[\frac{s}{m} - G_n(x'_s) \right], \quad (2)$$

$$D_{m,n} = \max(D^{+}_{m,n}, D^{-}_{m,n}), \quad (3)$$

где $x'_1 < x'_2 < \ldots < x'_m$ и $y'_1 < y'_2 < \ldots < y'_n$ — элементы первой $x_1, x_2, \ldots, x_m$ и второй $y_1, y_2, \ldots, y_n$ выборок, переставленные в порядке возрастания. Поскольку функции распределения $F(x)$ и $G(x)$ предполагаются непрерывными, то вероятность совпадения каких-либо выборочных значений равна нулю.

Пример 1. Пусть, как и в работе [2], даны две выборки. Первая содержит m = 12 элементов: 17; 22; 3; 5; 15; 2; 0; 7; 13; 97; 66; 14; вторая — n = 14 элементов: 47; 30; 2; 15; 1; 21; 25; 7; 44; 29; 33; 11; 6; 15. Проведем проверку однородности функций распределения двух выборок с помощью критерия Смирнова.

Переставим элементы первой и второй выборок в порядке возрастания: 0 < 2 < 3 < 5 < 7 < 13 < 14 < 15 < < 17 < 22 < 66 < 97 и 1 < 2 < 6 < 7 < 11 < 15 = 15 < 21 < < 25 < 29 < 30 < 33 < 44 < 47. Элементы второй выборки переставлены, точнее, в порядке неубывания, поскольку два элемента совпадают. С точки зрения теории вероятность совпадения двух элементов равна 0, но из-за неизбежных округлений она положительна. Поскольку совпадений мало (как внутри одной выборки, так и для элементов разных выборок), то использование теории, основанной на нулевой вероятности совпадения элементов выборок, является допустимым.

Расчет значений статистик $D^{+}_{m,n}$ и $D^{-}_{m,n}$ целесообразно проводить с помощью разработанных нами табл. 1 и 2 соответственно.

Беря максимум по столбцу **6** табл. 1, получаем, что $D^{+}_{m,n}$ = 0,167. Таков же максимум и по столбцу **9**, как и должно быть в соответствии с приведенным

выше равенством (1). В табл. 2 максимум по столбцу **6** равен 0,262, как и максимум по столбцу **9** [см. формулу (2)]. Согласно формуле (3) значение двухвыборочной статистики Смирнова

$$D_{m,n} = \max(0{,}167;\ 0{,}262) = 0{,}262.$$

В справочнике [4] (см. табл. 6.5а) приведены критические значения для двухвыборочной статистики Смирнова, соответствующие обычно используемым уровням значимости (табл. 3). Поскольку полученное по статистическим данным значение меньше критического для уровня значимости $\alpha = 0{,}1$, а потому и для всех остальных рассматриваемых уровней значимости, то нет оснований отклонять гипотезу однородности. Как и при использовании критерия Вилкоксона [2], эффект не обнаружен, нулевую гипотезу абсолютной однородности принимаем.

Разработаны алгоритмы и программы для ЭВМ, позволяющие рассчитывать точные распределения, процентные точки и достигаемый уровень значимости для двухвыборочной статистики Смирнова $D_{m,n}$, а также подробные таблицы (см., например, разработанную под нашим руководством методику [5], содержащую описание алгоритмов, тексты программ и подробные таблицы).

Однако у критерия Смирнова есть и недостатки. Его распределение сосредоточено в сравнительно небольшом числе точек. Ясно, что принимаемые этой статистикой значения пропорциональны величине $1/L$, где L — наименьшее общее кратное объемов выборок m и n. Поэтому функция распределения растет большими скачками. Для рассматриваемого примера L — наименьшее общее кратное 12 и 14, т.е. 84. Следовательно, принимаемые статистикой Смирнова величины входят в арифметическую прогрессию с ша-

Таблица 2. Данные для расчета значения статистики $D^{-}_{m,n}$

Номер элемента в объединенной выборке	Элементы выборок x	Номера выборок	$F_m(x)$	$\frac{r-1}{n}$	$F_m(x) - \frac{r-1}{n}$	$G_n(x)$	$\frac{s}{m}$	$\frac{s}{m} - G_n(x)$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	1	0			0	0,083	0,083
2	1	2	0,083	0	0,083	0		
3	2	1	0,083			0,071	0,167	0,096
4	2	2	0,083	0,071	0,012	0,071		
5	3	1	0,167			0,143	0,25	0,107
6	5	1	0,25			0,143	0,333	0,19
7	6	2	0,333	0,143	0,19	0,143		
8	7	1	0,333			0,214	0,417	0,203
9	7	2	0,333	0,214	0,119	0,214		
10	11	2	0,417	0,286	0,131	0,286		
11	13	1	0,417			0,357	0,5	0,143
12	14	1	0,5			0,357	0,583	0,226
13	15	1	0,583			0,357	0,667	0,31
14	15	2	0,583	0,357	0,226	0,357		
15	15	2	0,583	0,429	0,154	0,357		
16	17	1	0,667			0,5	0,75	0,25
17	21	2	0,75	0,5	0,25	0,5		
18	22	1	0,75			0,571	0,833	0,262
19	25	2	0,833	0,571	0,262	0,571		
20	29	2	0,833	0,643	0,19	0,643		
21	30	2	0,833	0,714	0,119	0,714		
22	33	2	0,833	0,786	0,047	0,786		
23	44	2	0,833	0,857	–0,024	0,857		
24	47	2	0,833	0,929	–0,096	0,929		
25	66	1	0,833			1,0	0,917	–0,083
26	97	1	0,917			1,0	1,0	0

Таблица 3. Критические значения и истинные уровни значимости для двухвыборочной статистики Смирнова ($m = 12$, $n = 14$)

Критические значения и истинный уровень значимости	Номинальный уровень значимости, %			
	10	5	2	1
Критическое значение (дробь)	39/84	43/84	47/84	52/84
Критическое значение (десятичное число)	0,464	0,512	0,559	0,619
Истинный уровень значимости	8,7	4,4	2,0	0,8

гом 1/84 = 0,012. Именно поэтому в сборнике [4] критические значения приведены в виде дроби с знаменателем $L = 84$.

Кроме того, не удается выдержать заданный уровень значимости. Реальный (другими словами, истинный) уровень значимости может значительно, даже в несколько раз отличаться от номинального (подробному обсуждению неклассического феномена существенного отличия реального уровня значимости от номинального посвящена работа [1]).

При больших объемах выборок можно воспользоваться доказанной Н. В. Смирновым в 1939 г. теоремой: в случае совпадения непрерывных функций распределения элементов двух независимых выборок

$$\lim_{m\to\infty,\ n\to\infty} P\left\{\sqrt{\frac{mn}{m+n}}D_{m,n} < y\right\} = K(y),$$

где

$$K(y) = \sum_{k=-\infty}^{\infty}(-1)^k \exp[-2k^2 y_2]—$$

функция распределения Колмогорова.

Поскольку [4] квантиль порядка 0,9 функции распределения Колмогорова равен 1,224, то критическое значение двухвыборочной статистики Смирнова $D_{m,n}$, соответствующее уровню значимости 10 %, при больших объемах выборок имеет вид

$$1{,}224\sqrt{\frac{m+n}{mn}}.$$

При $m = 12$, $n = 14$ эта формула дает 0,4815, в то время как точное значение равно 0,464 (см. табл. 3). Видим, что приближение удовлетворительное, т.е. рассматриваемые объемы выборок (более 10 элементов) можно считать большими. Для построения правил принятия решений на основе значений двухвыборочной статистики Смирнова, соответствующих другим уровням значимости, можно воспользоваться квантилями функции распределения Колмогорова, взятыми из справочника [4]:

Величина a	0,8	0,9	0,95	0,98	0,99
Квантиль порядка a	1,07275	1,22385	1,35810	1,51743	1,62762

Критерий типа ω² (Лемана – Розенблатта)

Статистика критерия типа ω² для проверки однородности двух независимых выборок имеет вид

$$A = \frac{mn}{m+n}\int_{-\infty}^{\infty}(F_m(x) - G_n(x))^2 \mathrm{d}H_{m+n}(x), \qquad (4)$$

где $H_{m+n}(x)$ — эмпирическая функция распределения,

Таблица 4. Данные для расчета значения статистики A Лемана – Розенблатта

Номер элемента в объединенной выборке	Элементы выборок x	Номера выборок	i	$r_i - i$	$(r_i - i)^2$	j	$s_j - j$	$(s_j - j)^2$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	1	1	0	0			
2	1	2				1	1	1
3	2	1	2	1	1			
4	2	2				2	2	4
5	3	1	3	2	4			
6	5	1	4	2	4			
7	6	2				3	4	16
8	7	1	5	3	9			
9	7	2				4	5	25
10	11	2				5	5	25
11	13	1	6	5	25			
12	14	1	7	5	25			
13	15	1	8	5	25			
14	15	2				6	8	64
15	15	2				7	8	64
16	17	1	9	7	49			
17	21	2				8	9	81
18	22	1	10	8	64			
19	25	2				9	10	100
20	29	2				10	100	100
21	30	2				11	10	100
22	33	2				12	10	100
23	44	2				13	10	100
24	47	2				14	10	100
25	66	1	11	14	196			
26	97	1	12	14	196			

построенная по объединенной выборке. Несложно увидеть, что

$$H_{m+n}(x) = \frac{m}{m+n} F_m(x) + \frac{n}{m+n} G_n(x).$$

Статистика A типа ω^2 была предложена Э. Леманом в 1951 г., изучена М. Розенблаттом в 1952 г., а затем и другими исследователями. Она зависит лишь от рангов элементов двух выборок в объединенной выборке. Пусть $x_1, x_2, ..., x_m$ — первая выборка, $x'_1 < x'_2 < ... < x'_m$ — соответствующий вариационный ряд; $y_1, y_2, ..., y_n$ — вторая выборка, $y'_1 < y'_2 < ... < y'_n$ — соответствующий ей вариационный ряд. Поскольку функции распределения независимых выборок непрерывны, то с вероятностью единица все выборочные значения различны, совпадения отсутствуют. Из формулы (4) получим [4]

$$A = \frac{1}{mn(m+n)} \left[m \sum_{i=1}^{m} (r_i - i)^2 + n \sum_{j=1}^{n} (s_j - j)^2 \right] - \frac{4mn-1}{6(m+n)},$$

где r_i — ранг x'_i и s_j — ранг y'_j в общем вариационном ряду, построенном по объединенной выборке.

Данные для расчета значения статистики A типа ω^2 (статистики Лемана – Розенблатта) представлены в разработанной нами табл. 4. Суммируя значения в столбце **6**, получаем

$$\sum_{i=1}^{12} (r_i - i)^2 = 598.$$

С помощью столбца **9** находим

$$\sum_{j=1}^{14} (s_j - j)^2 = 880.$$

Следовательно,

$$A = \frac{12 \cdot 598 + 14 \cdot 880}{12 \cdot 14 \cdot 26} - \frac{4 \cdot 12 \cdot 14 - 1}{6 \cdot 26} =$$

$$= \frac{7176 + 12320}{4368} - \frac{671}{156} = 4{,}4634 - 4{,}3013 = 0{,}1621.$$

Известно [6], что (в обозначениях [4])

$$\lim_{m \to inf,\ n \to \infty} P[A < x] = a_1(x),$$

где $a_1(x)$ — предельная функция распределения классической статистики ω^2 (Крамера – Мизеса – Смирнова), используемой для проверки согласия эмпирического распределения с заданным теоретическим.

Квантили функции распределения $a_1(x)$ приведены ниже.

Величина a	0,8	0,9	0,95	0,98	0,99
Квантиль порядка a	0,245	0,347	0,461	0,620	0,743

Известно [4, 6], что в случае статистики Лемана – Розенблатта предельным распределением можно пользоваться и для выборок умеренного объема (5 и 7, 6 и 7, 7 и 7, 8 и 8 и т.д.). Поскольку наблюдаемое значение $A = 0{,}1621$ меньше любого приведенного выше критического значения, то гипотезу однородности двух рассматриваемых выборок следует принять.

Для критерия типа ω^2 (Лемана – Розенблатта) нет выраженного эффекта различия между номинальными и реальными уровнями значимости. Поэтому для проверки абсолютной однородности функций распределения (гипотеза H_0) рекомендуем применять статистику A типа ω^2. Если методическое, табличное или программное обеспечение для статистики Лемана – Розенблатта отсутствует, можно использовать критерий Смирнова.

ЛИТЕРАТУРА

1. **Камень Ю. Э., Камень Я. Э., Орлов А. И.** Реальные и номинальные уровни значимости в задачах проверки статистических гипотез / Заводская лаборатория. 1986. Т. 52. № 12. С. 55 – 57.
2. **Орлов А. И.** Какие гипотезы можно проверять с помощью двухвыборочного критерия Вилкоксона? / Заводская лаборатория. Диагностика материалов. 1999. Т. 65. № 1. С. 51 – 55.
3. **Орлов А. И.** О проверке однородности двух независимых выборок / Заводская лаборатория. Диагностика материалов. 2003. Т. 69. № 1. С. 55 – 60.
4. **Большев Л. Н., Смирнов Н. В.** Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
5. Методика. Проверка однородности двух выборок параметров продукции при оценке ее технического уровня и качества. — М.: ВНИИ стандартизации, 1987. — 116 с.
6. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.