# Математические методы исследования
## Mathematical methods of investigation

DOI: 10.26896/1028-6861-2017-83-11-66-72 УДК (UDC) 519.28

# МОДЕЛЬ АНАЛИЗА СОВПАДЕНИЙ ПРИ РАСЧЕТЕ НЕПАРАМЕТРИЧЕСКИХ РАНГОВЫХ СТАТИСТИК

© **А. И. Орлов**

Институт высоких статистических технологий и эконометрики Московского государственного технического университета им. Н. Э. Баумана; Московский физико-технический институт, Москва, Россия; e-mail: prof-orlov@mail.ru

*Статья поступила 24 января 2017 г.*

Непараметрическая статистика — одна из точек роста современных математико-статистических методов исследования. В непараметрической статистике важное место занимают ранговые критерии, основанные на использовании рангов элементов выборок (результатов наблюдений), а не самих числовых значений элементов выборок. Ранги — это номера элементов выборок в соответствующих вариационных рядах, построенных путем перестановки результатов наблюдений в порядке неубывания. Распределения ранговых критериев получены в предположении, что функции распределения результатов наблюдений непрерывны. Из этого предположения следует, что вероятность совпадения значений случайных величин, образующих анализируемые выборки, равна нулю. Однако в реальных данных встречаются совпадения. Следовательно, неверно предположение, что функции распределения результатов наблюдений непрерывны, а потому известными теоремами о распределениях ранговых статистик, строго говоря, пользоваться нельзя. Тем не менее при небольшом числе совпадений обычно рекомендуют применять ранговые статистики, иногда вводя те или иные поправки. Таким образом, над классической математико-статистической теорией устанавливают дополнительную надстройку в целях учета совпадения данных. Естественно, возникает вопрос о степени обоснованности тех или иных методов учета совпадения данных расчета. Предлагаем вероятностно-статистическую модель, объясняющую появление совпадений и дающую алгоритмы анализа совпадений. Эта модель основана на предположении о появлении совпадений данных в результате «слипания» мало различающихся результатов наблюдений. Поэтому добавляем малую поправку к каждому элементу совпадающей группы результатов наблюдений и в результате получаем выборку без совпадений, для которой рассчитываем значение ранговой статистики. Рассмотрев различные варианты поправок, получаем «облако» значений ранговой статистики. Анализ этого «облака» позволяет получить статистические выводы. В качестве примера рассмотрен двухвыборочный критерий Вилкоксона.

**Ключевые слова:** непараметрическая статистика; ранговые критерии; совпадение данных; вероятностно-статистическая модель; алгоритм анализа совпадений; двухвыборочный критерий Вилкоксона.

# THE MODEL OF COINCIDENCE ANALYSIS IN THE CALCULATION OF NONPARAMETRIC RANK STATISTICS

© **A. I. Orlov**

*Submitted January 24, 2017.*

Nonparametric statistic is one of the points of growth of modern mathematical and statistical methods of research. In nonparametric statistics, rank criteria based on the use of the ranks of the sample elements (observation results), rather than numerical values of the sample elements themselves, take an important place. Ranks are the numbers of sample elements in the corresponding variation series, constructed by rearranging the results of observations in the order of nondecreasing. Distributions of rank criteria are obtained on the assumption of continuity of the distribution functions of the observation results, hence, the probability of coincidence of the values of the random variables forming the analyzed samples should be equal to zero. However, in actual data, there are coincidences. Consequently, the assumption of the continuity of the distribution functions of the observation results is incorrect and known theorems on the distribution of rank statistics, strictly speaking, are not applicable. However, with a small number of coincidences, ranks statistics can be recommended for use, albeit with some corrections. Thus, an additional superstructure is mounted on the classical mathematical-statistical theory to take into account the coincidence of the data. Naturally,

the validity of different methods used for accounting the coincidence of calculation data should be considered. We propose a probabilistic-statistical model that explains the occurrence of the coincidences and provides algorithms for their analysis. This model is based on the assumption that data coincidences appears as a result of "sticking together" of the slightly different observation results. We propose to introduce small corrections into each elements of the coincident group of observation results and thus to obtain a sample without coincidences and calculate the value of rank statistics. Having considered various variants of amendments, we obtain a "cloud" of values of rank statistics. Analysis of this "cloud" allows us to obtain statistical conclusions. Two-sample Wilcoxon test is considered as an example.

**Keywords:** nonparametric statistics; rank criteria; data coincidence; probabilistic-statistical model; coincidence analysis algorithm; Wilcoxon two-sample test.

**Н**епараметрическая статистика — одна из точек роста современных математико-статистических методов исследования [1]. В непараметрической статистике важное место занимают ранговые критерии, основанные на использовании рангов наблюдений, а не самих результатов измерений, наблюдений, испытаний, опытов, анализов, обследований (одним словом, — данных). Ранги — это номера результатов наблюдений в соответствующих вариационных рядах, построенных путем перестановки результатов наблюдений в порядке неубывания.

Непараметрическая статистика началась с критерия Колмогорова, предназначенного для проверки согласия выборочного распределения с заданным теоретическим [3]. Другими примерами ранговых критериев являются критерий Вилкоксона [4, 5] и другие критерии проверки однородности двух независимых выборок, прежде всего состоятельные критерии Смирнова и Лемана – Розенблатта [6], а также критерий Орлова типа омега-квадрат [7, 8] для проверки однородности связанных выборок (симметрии распределения относительно нуля).

Как известно, распределения ранговых критериев получены в предположении, что функции распределения результатов наблюдений непрерывны. Из этого предположения следует, что вероятность совпадения значений случайных величин, образующих анализируемые выборки, равна нулю.

Однако в реальных данных встречаются совпадения. Следовательно, неверно предположение, что функции распределения результатов наблюдений непрерывны, а потому известными теоремами о распределениях ранговых статистик пользоваться нельзя.

Тем не менее при небольшом числе совпадений обычно рекомендуют применять ранговые статистики, иногда вводя те или иные поправки [9]. Таким образом, над классической математико-статистической теорией устанавливают дополнительную надстройку в целях учета совпадения данных. Естественно, возникает вопрос о степени обоснованности рекомендуемых методов расчета.

Строго говоря, отклонения от классической теории имеют место и при отсутствии совпадения данных. Дело в том, что представленные для анализа данные записаны с определенным числом значащих цифр, т.е. принадлежат некоторому множеству из конечного числа элементов. Вероятность же того, что значение случайной величины с непрерывной функцией распределения принадлежит этому конечному множеству, равна нулю. С вероятностью единица ее значение должно описываться действительным числом с бесконечным числом значащих цифр, что возможно только в теории. Таким образом, обрабатываемые данные с вероятностью единица не соответствуют классической математико-статистической теории [10].

## Моделирование как фундамент разработки статистических методов

В соответствии с современными подходами к математическим методам исследования необходимо, чтобы сначала была описана математическая (вероятностно-статистическая) модель порождения данных (тем самым сделан переход от реального мира внутрь математической конструкции) и только после этого проведены расчеты (внутри математики). Затем переходят к заключительному этапу исследования — математические результаты интерпретируются в терминах предметной области.

Типовая ошибка слабо подготовленных исследователей — игнорирование математической (вероятностно-статистической) модели. Они сразу начинают строить здание без фундамента — проводить расчеты.

Модели могут быть разные. Важно, чтобы они были, причем обоснованные.

Для описания результатов измерений целесообразно использовать основную модель статистики интервальных данных [11], согласно которой эти измерения отягощены ошибками.

Точнее, пусть сущность реального явления описывается выборкой $x_1, x_2, ..., x_n$. В вероятностной теории математической статистики выборка моделируется набором независимых в совокупности одинаково распределенных случайных величин. Однако беспристрастный и тщательный анализ подавляющего большинства реальных задач показывает, что статистику известна отнюдь не выборка $x_1, x_2, ..., x_n$, а величины

$$y_j = x_j + \varepsilon_j,\ j = 1, 2, ..., n,$$

где $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ — некоторые погрешности измерений, наблюдений, анализов, опытов, исследований (например, инструментальные ошибки). Одна из причин появления погрешностей — запись результатов наблюдений с конечным числом значащих цифр.

Если ограничения на погрешности имеют вид

$$|\varepsilon_i| \le \Delta, \ \ i = 1, 2, ..., n,$$

причем $\Delta$ мало, то для различающихся значений из множества чисел $y_1$, $y_2$, ..., $y_n$ их ранги совпадают с рангами соответствующих чисел из множества $x_1$, $x_2$, ..., $x_n$ (точнее, вероятность такого совпадения стремится к единице при $\Delta \to 0$). Поэтому достаточно рассмотреть вероятностно-статистическую модель для совпадающих значений в анализируемых выборках.

Основная идея состоит в том, что совокупность совпадающих значений $y_1 = y_2 = ... = y_k = a$ получена из набора $x_j = y_j - \varepsilon_j$, $j = 1, 2, ...$, при некоторых малых отклонениях $\varepsilon_j$, $j = 1, 2, ..., k$. Одна точка $a$ соответствует «облаку» из $k$ точек. Соответственно получаем «облако» упорядочений, состоящее из $k!$ ранжировок, каждая из которых реализуется с вероятностью $1/k!$. Каждое упорядочение порождает соответствующее значение ранговой статистики, а все ранжировки — «облако» таких значений. При наличии нескольких групп совпадающих значений упорядочения комбинируются и число элементов итогового «облака» равно произведению факториалов, соответствующих отдельным облакам. Поскольку некоторые значения в облаке иногда совпадают, то число различных элементов в итоговом «облаке» может быть меньше указанного.

Анализ описанных «облаков» проводится разными способами. Можно рассчитать «исправленное» значение ранговой статистики, усреднив элементы «облака» в соответствии с вероятностями их появления. Нам представляется более практичным подход на основе теории устойчивости статистических процедур [12], в простейшем случае сводящийся к расчету размаха возможных значений ранговой статистики (т.е. значений, входящих в «облако»).

Не стремясь к построению достаточно общей, а потому математизированной теории, продемонстрируем сказанное на примере широко распространенного, хорошо изученного и активно используемого рангового критерия.

## Анализ совпадений для двухвыборочного критерия Вилкоксона

Рассмотрим численный пример. Пусть 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. Опишем проверку однородности функций распределения двух независимых выборок с помощью критерия Вилкоксона. Первым шагом является построение общего вариационного ряда для элементов двух выборок (табл. 1).

Хотя с точки зрения теории математической статистики вероятность совпадения двух элементов выборок из непрерывных распределений равна нулю, в реальных выборках статистических данных совпадения встречаются. Так, в рассматриваемых выборках (см. табл. 1) два раза повторяется величина 2, два раза — 7 и три раза — 15. В таких случаях в целях расчета «исправленного» значения ранговой статистики говорят о наличии «связанных рангов», а соответствующим совпадающим величинам приписывают среднее арифметическое тех рангов, которые они занимают. Так, результаты измерений 2 и 2 занимают в объединенной выборке места 3 и 4, поэтому им приписывается ранг $(3 + 4)/2 = 3{,}5$, величины 7 и 7 — места 8 и 9, поэтому им приписывается ранг $(8 + 9)/2 = 8{,}5$, величины 15, 15 и 15 — места 13, 14 и 15, поэтому им приписывается ранг $(13 + 14 + 15)/3 = 14$.

Следующий шаг — подсчет значения статистики Вилкоксона, т.е. суммы рангов элементов первой выборки

$$S = R_1 + R_2 + ... + R_m = 1 + 3{,}5 + 5 + 6 + 8{,}5 + 11 + 12 +$$

$$+ 14 + 16 + 18 + 25 + 26 = 146.$$

Подсчитаем также сумму рангов элементов второй выборки

$$S_1 = 2 + 3{,}5 + 7 + 8{,}5 + 10 + 14 + 14 + 17 + 19 + 20 +$$

$$+ 21 + 22 + 23 + 24 = 205.$$

Величина $S_1$ может быть использована для контроля вычислений. Дело в том, что суммы рангов элементов первой выборки $S$ и второй выборки $S_1$ вместе составляют сумму рангов объединенной выборки, т.е. сумму всех натуральных чисел от 1 до $m + n$. Следовательно,

$$S + S_1 = (m + n)(m + n + 1)/2 =$$

$$= (12 + 14)(12 + 14 + 1)/2 = 351.$$

**Таблица 1.** Общий вариационный ряд для элементов двух выборок

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ранги | 1 | 2 | 3,5 | 3,5 | 5 | 6 | 7 | 8,5 | 8,5 | 10 | 11 | 12 | 14 |
| Элементы выборок | 0 | 1 | 2 | 2 | 3 | 5 | 6 | 7 | 7 | 11 | 13 | 14 | 15 |
| Номера выборок | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 |
| Ранги | 14 | 14 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| Элементы выборок | 15 | 15 | 17 | 21 | 22 | 25 | 29 | 30 | 33 | 44 | 47 | 66 | 97 |
| Номера выборок | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |

В соответствии с ранее проведенными расчетами $S + S_1 = 146 + 205 = 351$. Необходимое условие правильности расчетов выполнено. Ясно, что справедливость этого условия не гарантирует правильности расчетов.

На практике проверять гипотезу однородности двух независимых выборок можно с помощью таблиц критических значений статистики Вилкоксона $S$. Однако в них ограничен диапазон возможных значений объемов выборок. Поэтому целесообразно строить алгоритм принятия решения на основе центрированной нормированной статистики Вилкоксона $T$. Перейдем к ее расчету. Согласно [4, 5]

$$T = \frac{S - M(S)}{\sqrt{D(S)}},$$

где математическое ожидание статистики $S$

$$M(S) = m(m + n + 1)/2,$$

а дисперсия статистики $S$

$$D(S) = mn(m + n + 1)/12.$$

Исходим из асимптотической нормальности статистики $T$: при безграничном росте объемов выборок ($m \to \infty$, $n \to \infty$) функция распределения $P(T < x)$ статистики $T$ стремится к функции $\Phi(x)$ с математическим ожиданием 0 и дисперсией 1 [13]. Асимптотическая нормальность статистики $T$ позволяет задавать критические значения критерия проверки однородности двух независимых выборок, построенного на основе этой статистики, с помощью квантилей стандартного нормального распределения.

Согласно сказанному выше

$$M(S) = 12(12 + 14 + 1)/2 = 162,$$

$$D(S) = 12 \cdot 14(12 + 14 + 1)/12 = 378.$$

Следовательно,

$$T = (S - 162)(378)^{-1/2} = (146 - 162)/19{,}44 = -0{,}82.$$

Поскольку $|T| < 1{,}96$, то гипотеза однородности принимается на уровне значимости 0,05.

Что будет, если поменять выборки местами и вторую назвать первой? Тогда вместо $S$ надо рассматривать $S_1$. Имеем

$$M(S_1) = 14(12 + 14 + 1)/2 = 189,$$

$$D(S) = D(S_1) = 378,$$

$$T_1 = (S_1 - 189)(378)^{-1/2} = (205 - 162)/19{,}44 = 0{,}82.$$

Таким образом, значения статистики критерия отличаются только знаком (можно показать, что это утверждение верно всегда). Поскольку в правиле принятия решения используется только абсолютная величина статистики, то принимаемое решение не зависит от того, какую выборку считаем первой, а какую второй. Для уменьшения объема таблиц критических значений ранее было принято считать первой выборку меньшего объема.

Построим «облако» упорядочений, соответствующее введенной выше модели анализа совпадающих значений. Поскольку в выборку входят целые числа, то достаточно принять, что максимальная абсолютная погрешность не превосходит $\Delta < 0{,}5$. Результаты построения «облака» упорядочений приведены в табл. 2. В ней выделены три группы совпадающих значений, для которых ранее в табл. 1 были введены связанные ранги. Значение 2 имеется как в первой выборке, так и во второй. При построении «облака» меньшим может быть либо реальное значение из первой выборки (упорядочения 1 – 6), либо реальное значение из второй выборки (упорядочения 7 – 12) — всего две возможности. Аналогична ситуация со значением 7. Оно имеется как в первой выборке, так и во второй. При по-

**Таблица 2.** «Облако» упорядочений для совпадающих данных табл. 1.

| | Связанные ранги | | | | | | | Статистика $S$ |
|---|---|---|---|---|---|---|---|---|
| | 3,5 | 3,5 | 8,5 | 8,5 | 14 | 14 | 14 | |
| Упорядочение 1 | 3 | 4 | 8 | 9 | 13 | 14 | 15 | 144 |
| Упорядочение 2 | 3 | 4 | 8 | 9 | 14 | 13 | 15 | 145 |
| Упорядочение 3 | 3 | 4 | 8 | 9 | 15 | 13 | 14 | 146 |
| Упорядочение 4 | 3 | 4 | 9 | 8 | 13 | 14 | 15 | 145 |
| Упорядочение 5 | 3 | 4 | 9 | 8 | 14 | 13 | 15 | 146 |
| Упорядочение 6 | 3 | 4 | 9 | 9 | 15 | 13 | 14 | 147 |
| Упорядочение 7 | 4 | 3 | 8 | 9 | 13 | 14 | 15 | 145 |
| Упорядочение 8 | 4 | 3 | 8 | 9 | 14 | 13 | 15 | 146 |
| Упорядочение 9 | 4 | 3 | 8 | 9 | 15 | 13 | 14 | 147 |
| Упорядочение 10 | 4 | 3 | 9 | 8 | 13 | 14 | 15 | 146 |
| Упорядочение 11 | 4 | 3 | 9 | 8 | 14 | 13 | 15 | 147 |
| Упорядочение 12 | 4 | 3 | 9 | 9 | 15 | 13 | 14 | 148 |
| Элементы выборок | 2 | 2 | 7 | 7 | 15 | 15 | 15 | |
| Номера выборок | 1 | 2 | 1 | 2 | 1 | 2 | 2 | |

строении «облака» меньшим может быть либо реальное значение из первой выборки (упорядочения 1 – 3 и 7 – 9), либо реальное значение из второй выборки (упорядочения 4 – 6 и 10 – 12) — всего две возможности. Несколько сложнее со значением 15. Оно один раз встречается в первой выборке и два раза — во второй. Следовательно, реальное упорядочение состоит из трех элементов. При построении «облака» реальное значение из первой выборки может стоять либо на первом месте (упорядочения 1, 4, 7, 10), либо на втором (упорядочения 2, 5, 8, 11), либо на третьем месте (упорядочения 3, 6, 9, 12), остальные два места занимают реальные значения из второй выборки — всего три возможности. Таким образом, облако состоит из $2 \cdot 2 \cdot 3 = 12$ упорядочений.

Подсчитаем значения статистики Вилкоксона, т.е. сумму рангов элементов первой выборки. Как показано выше, при использовании взвешенных рангов $S = S_0 = 146$.

Для упорядочения 1

$$S = S_1 = 1 + 3 + 5 + 6 + 8 + 11 + 12 + 13 + 16 + 18 + 25 + 26 = S_0 - 0{,}5 - 0{,}5 - 1 = 144;$$

для упорядочения 2

$$S = S_2 = 1 + 3 + 5 + 6 + 8 + 11 + 12 + 14 + 16 + 18 + 25 + 26 = S_0 - 0{,}5 - 0{,}5 = 145;$$

для упорядочения 3

$$S = S_3 = 1 + 3 + 5 + 6 + 8 + 11 + 12 + 15 + 16 + 18 + 25 + 26 = S_0 - 0{,}5 - 0{,}5 + 1 = 146;$$

для упорядочения 4

$$S = S_4 = 1 + 3 + 5 + 6 + 9 + 11 + 12 + 13 + 16 + 18 + 25 + 26 = S_0 - 0{,}5 + 0{,}5 - 1 = 145;$$

для упорядочения 5

$$S = S_5 = 1 + 3 + 5 + 6 + 9 + 11 + 12 + 14 + 16 + 18 + 25 + 26 = S_0 - 0{,}5 + 0{,}5 = 146;$$

для упорядочения 6

$$S = S_6 = 1 + 3 + 5 + 6 + 9 + 11 + 12 + 15 + 16 + 18 + 25 + 26 = S_0 - 0{,}5 + 0{,}5 + 1 = 147;$$

для упорядочения 7

$$S = S_7 = 1 + 4 + 5 + 6 + 8 + 11 + 12 + 13 + 16 + 18 + 25 + 26 = S_0 + 0{,}5 - 0{,}5 - 1 = 145;$$

для упорядочения 8

$$S = S_8 = 1 + 4 + 5 + 6 + 8 + 11 + 12 + 14 + 16 + 18 + 25 + 26 = S_0 + 0{,}5 - 0{,}5 = 146;$$

для упорядочения 9

$$S = S_9 = 1 + 4 + 5 + 6 + 8 + 11 + 12 + 15 + 16 + 18 + 25 + 26 = S_0 + 0{,}5 - 0{,}5 + 1 = 147;$$

для упорядочения 10

$$S = S_{10} = 1 + 4 + 5 + 6 + 9 + 11 + 12 + 13 + 16 + 18 + 25 + 26 = S_0 + 0{,}5 + 0{,}5 - 1 = 146;$$

для упорядочения 11

$$S = S_{11} = 1 + 4 + 5 + 6 + 9 + 11 + 12 + 14 + 16 + 18 + 25 + 26 = S_0 + 0{,}5 + 0{,}5 = 147;$$

для упорядочения 12

$$S = S_{12} = 1 + 4 + 5 + 6 + 9 + 11 + 12 + 15 + 16 + 18 + 25 + 26 = S_0 + 0{,}5 + 0{,}5 + 1 = 148.$$

Поскольку все 12 упорядочений равновероятны, то распределение значений статистики Вилкоксона в «облаке» следующее.

| | | | | | |
|---|---|---|---|---|---|
| Значения . . . . . | 144 | 145 | 146 | 147 | 148 |
| Вероятности . . . . . | 1/12 | 3/12 | 4/12 | 3/12 | 1/12 |

Таким образом, анализ совпадений позволил перейти от точечного значения статистики Вилкоксона к интервалу [144, 148]. Разброс небольшой, влияние на выводы мало. Действительно, значение центрированной и нормированной статистики Вилкоксона $T$ лежит в интервале $[T_1, T_2]$, где

$$T_1 = (144 - 162)/19{,}44 = -0{,}93,$$

$$T_1 = (148 - 162)/19{,}44 = -0{,}72.$$

Ввиду небольшого разброса вывод о принятии нулевой гипотезы об однородности сохраняется.

Таким образом, работа посвящена анализу особенностей применения критериев непараметрической статистики в условиях несоблюдения важного теоретического предположения о том, что функции распределения результатов наблюдений непрерывны. Для реальных выборок такое предположение редко выполняется и практически всегда имеются совпадающие значения. Это заставляет ряд исследователей вводить специальные поправки, учитывающие наличие «связанных рангов». Автор предлагает другой подход — на основе теории устойчивости статистических процедур рассчитать интервал возможных значений ранговой статистики. Для пояснения подхода рассматри-

вается пример, в котором проводится проверка однородности функций распределения двух независимых выборок с помощью двухвыборочного критерия Вилкоксона (Манна – Уитни). Вместо обычно применяемой точечной оценки в работе получен интервал значений статистики Вилкоксона.

Предлагается следующий алгоритм анализа совпадений при расчете непараметрических ранговых статистик.

1. В объединенной выборке выделить группы совпадающих значений.

2. На основе каждой группы совпадающих значений построить совокупность групп элементов объединенной выборки (названную выше «облаком»), добавляя к каждому входящему в группу элементу выборки малое положительное или отрицательное число.

*Комментарий к п. 2 алгоритма.* Основная идея состоит в том, что совокупность представленных для статистического анализа совпадающих значений $y_1 = y_2 = ... = y_k = a$ получена (при связанных с процессом измерения округлениях) из набора $x_j = y_j - \varepsilon_j$, $j = 1, 2, ..., k$, при некоторых малых отклонениях $\varepsilon_j$, $j = 1, 2, ..., k$, среди которых нет совпадающих. Одна точка $a$ соответствует «облаку» из $k$ различных точек $x_j = a - \varepsilon_j, j = 1, 2, ..., k$.

3. Упорядочить элементы «облака».

*Комментарий к п. 3 алгоритма.* Для группы совпадающих значений $y_1 = y_2 = ... = y_k = a$ получаем «облако» упорядочений, состоящее из $k!$ ранжировок, каждая из которых реализуется с вероятностью $1/k!$.

4. Построить «облака» для всех групп совпадающих значений и скомбинировать их в единое «облако» для объединенной выборки.

*Комментарий к п. 4 алгоритма.* При наличии нескольких групп совпадающих значений упорядочения комбинируются и число элементов итогового «облака» равно произведению факториалов, соответствующих отдельным облакам. Итогом пп. 1 – 4 является «размножение» единой исходной объединенной выборки на некоторое число похожих, отличающихся введением того или иного упорядочения совпадающих значений. Другими словами, основная идея алгоритма настоящей работы лежит в общем русле использования различных процедур «размножения выборки» [14].

5. Рассчитать значение ранговой статистики для каждой выборки из единого «облака».

*Комментарий к п. 5 алгоритма.* Каждое упорядочение порождает соответствующее значение ранговой статистики, а все ранжировки — «облако» таких значений. Поскольку некоторые значения в «облаке значений» могут совпадать, то число различных элементов в итоговом «облаке» может быть меньше, чем произведение факториалов, соответствующих отдельным облакам.

6. Проанализировать распределение «облака значений» ранговой статистики и оценить влияние совпадений элементов выборок на статистические выводы.

*Комментарий к п. 6 алгоритма.* Анализ описанных «облаков» может проводиться разными способами. Один вариант — рассчитать «исправленное» значение ранговой статистики, усреднив элементы «облака» в соответствии с вероятностями их появления. Нам представляется более практичным подход на основе теории устойчивости статистических процедур [12], в простейшем случае сводящийся к расчету размаха возможных значений ранговой статистики (т.е. значений, входящих в «облако»).

Представляется ясным, что разработанный подход может быть успешно применен для различных непараметрических ранговых статистик при анализе совпадений анализируемых данных.

## ЛИТЕРАТУРА


1. **Орлов А. И.** Точки роста статистических методов / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 103. С. 136 – 162.
2. **Орлов А. И.** Структура непараметрической статистики (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 7. С. 62 – 72.
3. **Орлов А. И.** Непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат и ошибки при их применении / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 97. С. 32 – 45.
4. **Орлов А. И.** Какие гипотезы можно проверять с помощью двухвыборочного критерия Вилкоксона? / Заводская лаборатория. Диагностика материалов. 1999. Т. 65. № 1. С. 51 – 55.
5. **Орлов А. И.** Двухвыборочный критерий Вилкоксона — анализ двух мифов / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 104. С. 91 – 111.
6. **Орлов А. И.** Состоятельные критерии проверки абсолютной однородности независимых выборок / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 11. С. 66 – 70.
7. **Орлов А. И.** Методы проверки однородности связанных выборок / Заводская лаборатория. Диагностика материалов. 2004. Т. 70. № 7. С. 57 – 61.
8. **Орлов А. И.** О проверке однородности связанных выборок / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. № 123. С. 708 – 726.
9. **Холлендер М., Вулф Д. А.** Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
10. **Орлов А. И.** О методологии статистических методов / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 104. С. 53 – 80.
11. **Орлов А. И.** Статистика интервальных данных (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 3. С. 61 – 69.
12. **Орлов А. И.** Устойчивые математические методы и модели / Заводская лаборатория. Диагностика материалов. 2010. Т. 76. № 3. С. 59 – 67.
13. **Гаек Я., Шидак З.** Теория ранговых критериев. — М.: Наука, 1971. — 376 с.
14. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.


## REFERENCES


1. **Orlov A. I.** The growth points of statistical methods / Politem. Set. Élektron. Nauch. Zh. Kuban. Gos. Agrarn. Univ. 2014. N 103. P. 136 – 162 [in Russian].

2. **Orlov A. I.** Structure of nonparametric statistics (generalizing paper) / Zavod. Lab. Diagn. Mater. 2015. Vol. 81. N 7. P. 62 – 72 [in Russian].
3. **Orlov A. I.** Nonparametric goodness-of-fit Kolmogorov, Smirnov, omega-square tests and the errors in their application / Politem. Set. Élektron. Nauch. Zh. Kuban. Gos. Agrarn. Univ. 2014. N 97. P. 32 – 45 [in Russian].
4. **Orlov A. I.** What hypothesis can be verified using the two-sample Wilcoxon test? / Zavod. Lab. Diagn. Mater. 1999. Vol. 65. N 1. P. 51 – 55 [in Russian].
5. **Orlov A. I.** Two-sample Wilcoxon test — analysis of two myths / Politem. Set. Élektron. Nauch. Zh. Kuban. Gos. Agrarn. Univ. 2014. N 104. P. 91 – 111 [in Russian].
6. **Orlov A. I.** Consistent tests of absolute homogeneity for independent samples / Zavod. Lab. Diagn. Mater. 2012. Vol. 78. N 11. P. 66 – 70 [in Russian].
7. **Orlov A. I.** Methods for testing the homogeneity of the paired samples / Zavod. Lab. Diagn. Mater. 2004. Vol. 70. N 7. P. 57 – 61 [in Russian].
8. **Orlov A. I.** Testing of homogeneity of the paired samples / Politem. Set. Élektron. Nauch. Zh. Kuban. Gos. Agrarn. Univ. 2016. N 123. P. 708 – 726 [in Russian].
9. **Hollander M., Wolfe D. A., Chicken E.** Nonparametric Statistical Methods. Third Edition. — Hoboken, New Jersey: John Wiley & Sons, Inc., 2014. — 828 p.
10. **Orlov A. I.** About the methodology of statistical methods / Politem. Set. Élektron. Nauch. Zh. Kuban. Gos. Agrarn. Univ. 2014. N 104. P. 53 – 80 [in Russian].
11. **Orlov A. I.** Statistics of interval data (generalizing paper) / Zavod. Lab. Diagn. Mater. 2015. Vol. 81. N 3. P. 61 – 69 [in Russian].
12. **Orlov A. I.** Stable mathematical methods and models / Zavod. Lab. Diagn. Mater. 2010. Vol. 76. N 3. P. 59 – 67 [in Russian].
13. **Hajek Ja., Sidak Zb.** Theory of rank tests. — Prague: Academia. Publishing house of the Czechoslovak academy of sciences, 1967. — 376 p.
14. **Orlov A. I.** Applied statistics. — Moscow: Ékzamen, 2006. — 671 p. [in Russian].