

УДК 330.5 : 519.2

UDC 330.5 : 519.2

5.2.2. Математические, статистические и инструментальные методы в экономике (физико-математические науки, экономические науки)

5.2.2. Mathematical, statistical and instrumental methods of economics (physical and mathematical sciences, economic sciences)

НОВЫЙ ПОДХОД К АНАЛИЗУ СТАТИСТИЧЕСКИХ ДАННЫХ С СОВПАДЕНИЯМИ

A NEW APPROACH TO ANALYZING STATISTICAL DATA WITH COINCIDENCES

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н.,
профессор

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci.,
professor

РИНЦ SPIN-код: 4342-4994

RSCI SPIN-code: 4342-4994

prof-orlov@mail.ru

*Московский государственный технический
университет им. Н.Э. Баумана, Россия, 105005,
Москва, 2-я Бауманская ул., 5*

*Bauman Moscow State Technical University, Moscow,
Russia*

Непараметрическая статистика - одна из точек роста современных математико-статистических методов исследования. В ней 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. В теории математической статистики распределения ранговых критериев получены в предположении, что функции распределения результатов наблюдений непрерывны. Следовательно, вероятность совпадения значений случайных величин, образующих анализируемые выборки, равна 0. Однако в реальных данных встречаются совпадения. Таким образом, 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, вводя те или иные поправки, например, связанные ранги. Над классической математико-статистической теорией устанавливают дополнительную надстройку с целью учета совпадения данных. Естественно, возникает вопрос о степени обоснованности тех или иных методов учета совпадения данных расчета. Мы 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. Мы

Nonparametric statistics is one of the growth areas of modern mathematical and statistical research methods. Rank tests, based on the ranks of sample elements (observation results), rather than the numerical values of the sample elements themselves, occupy an important place in this field. Ranks are the numbers of sample elements in the corresponding variation series, constructed by rearranging the observation results in non-decreasing order. In the theory of mathematical statistics, the distributions of rank tests are obtained under the assumption that the distribution functions of the observation results are continuous. Consequently, the probability of coincidence of the values of random variables that form the analyzed samples is equal to 0. However, in real data, there are coincidences. Therefore, it is incorrect to assume that the distribution functions of the observation results are continuous, and therefore, strictly speaking, the known theorems about the distributions of rank statistics cannot be used. Nevertheless, when there are a small number of coincidences, it is usually recommended to use rank statistics by introducing certain corrections, such as related ranks. An additional layer is added to the classical mathematical and statistical theory to account for data coincidences. Naturally, the question arises as to the degree of validity of certain methods of accounting for the coincidence of calculation data. We propose a probabilistic-statistical model that explains the occurrence of coincidences and provides algorithms for analyzing coincidences. This model is based on the assumption that data coincidences occur as a result of the "clumping" of observations that differ little. We propose adding a small correction to each element of a matching group of observations to obtain a match-free sample for which to calculate the rank statistic. A model for generating matched data has been developed. A procedure for sample propagation

предлагаем добавить малую поправку к каждому элементу совпадающей группы результатов наблюдений с целью получить выборку без совпадений, для которой рассчитать значение ранговой статистики. Разработаны модель порождения данных с совпадениями. Предложена процедура размножения выборок. Рассмотрев различные варианты поправок, получаем «облако» значений ранговой статистики. Анализ этого «облака» позволяет получить обоснованные статистические выводы. В качестве примеров рассмотрены двухвыборочный критерий Вилкоксона для проверки однородности независимых выборок и критерий знаков для проверки однородности связанных выборок

is proposed. By considering various correction options, we obtain a "cloud" of rank statistic values. Analysis of this "cloud" allows for sound statistical inference. Examples include the Wilcoxon two-sample test for testing the homogeneity of independent samples and the sign test for testing the homogeneity of related samples

Ключевые слова: НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА, РАНГОВЫЕ КРИТЕРИИ, СОВПАДЕНИЕ ДАННЫХ, ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ, АЛГОРИТМ АНАЛИЗА СОВПАДЕНИЙ, ДВУХВЫБОРОЧНЫЙ КРИТЕРИЙ ВИЛКОКСОНА, КРИТЕРИЙ ЗНАКОВ

Keywords: NONPARAMETRIC STATISTICS, RANK CRITERIA, DATA COINCIDENCE, PROBABILITY-STATISTICAL MODEL, COINCIDENCE ANALYSIS ALGORITHM, TWO-SAMPLE WILCOXON TEST, SIGN TEST

<http://dx.doi.org/10.21515/1990-4665-221-020>

Введение

В соответствии с современной парадигмой математических методов исследований для анализа числовых статистических данных необходимо применять методы непараметрической статистики, в которой нет нереалистических предположений о том, что элементы выборки имеют распределение, входящее в то или иное параметрическое семейство. Так, хорошо известно, что распределения реальных данных, как правило, не являются нормальными [1, разд. 5.1].

В математической статистике базовая модель порождения данных – выборка, т.е. совокупность независимых одинаково распределенных случайных величин. Выборка применяется для описания результатов измерений, наблюдений, испытаний, опытов, анализов, обследований (короче – числовых статистических данных, которые рассматриваются как элементы выборки). Возможность применения этой модели порождения данных должна быть обоснована. Обычно такое обоснование основано на

анализе процедур получения статистических данных в соответствующей прикладной области, т.е. находится вне математической статистики. Анализ таких процедур может привести к выводу о необходимости использования других моделей порождения данных, например, потребовать отказа от предположения независимости элементов выборки и перехода к введению коррелированных величин. Так может обстоять дело при изучении временного ряда.

Непараметрическая статистика - одна из точек роста современных математико-статистических методов исследования [1, гл. 13]. В непараметрической статистике важное место занимают ранговые критерии, основанные на использовании не самих элементов выборки, а только их рангов. Ранги — это номера результатов наблюдений в соответствующих вариационных рядах, построенных путём перестановки результатов рассматриваемых наблюдений в порядке неубывания.

Непараметрическая статистика как самостоятельная область статистической науки началась с критерия Колмогорова, предназначенного для проверки согласия выборочного распределения с заданным теоретическим. К настоящему времени она является весьма развитой. Так, таблицам непараметрической статистики посвящён обширный раздел VI фундаментального справочника Л.Н. Большева и Н.В. Смирнова [2], содержащий развёрнутые комментарии.

В рассматриваемой области статистической науки важное место занимают ранговые критерии, являющиеся функциями от рангов элементов рассматриваемых выборок. Примерами ранговых критериев являются критерий Вилкоксона и другие критерии проверки однородности двух независимых выборок, прежде всего состоятельные критерии Смирнова и Лемана-Розенблатта, а также критерий знаков и критерий типа омега-квадрат для проверки однородности связанных выборок (т.е.

симметрии распределения разностей элементов связанных элементов относительно 0) [1, гл.8].

Первый шаг при построении любого статистического критерия – выбор модели порождения данных. Как уже говорилось, под выборкой обычно понимается совокупность независимых одинаково распределённых случайных величин. Важно, что теория ранговых критериев исходит из предположения, что функции распределения элементов рассматриваемых выборок непрерывны (см., например, [3]). Распределения рассматриваемых в математической статистике ранговых критериев найдены на основе этого математического предположения. Отметим, что из него следует также, что вероятность совпадения значений каких-либо двух случайных величин, входящих в анализируемые выборки, равна 0.

Однако в реальных данных, которые анализируют специалисты различных областей науки и практики, ситуация иная. Как хорошо известно, достаточно часто встречаются совпадения элементов выборок. Следовательно, неверно предположение, что функции распределения анализируемых результатов наблюдений непрерывны. Следовательно, на известные теоремы о распределениях ранговых статистик, строго говоря, нельзя опираться при анализе реальных данных.

В чём причина этого явления? В некоторых ситуациях используемая модель порождения данных основана на том, что распределения элементов выборки имеют дискретную составляющую, т.е. «атомы» - числа, для которых вероятность совпадения с ними элементов выборки положительна. Например, функция распределения случайного ущерба имеет выраженный «атом» в 0, поскольку рискованная ситуация осуществляется не всегда, а лишь в отдельных случаях.

Однако более важным (и более часто встречающимся) представляется отличие математических чисел, используемых в математических теориях, от прагматических чисел, которые применяются

при практических расчётах. Дело в том, что представленные для анализа данные на практике всегда записаны с определённым числом значащих цифр. Следовательно, они принадлежат некоторому множеству из конечного числа элементов. Как вытекает из теории вероятностей, случайное событие, состоящее в том, что значение случайной величины с непрерывной функцией распределения принадлежит какому-либо конечному множеству, имеет вероятность 0. Следовательно, с вероятностью 1 ее значение должно описываться действительным числом с бесконечным числом значащих цифр. Это возможно только в теории, поскольку все используемые на практике числа имеют конечное число значащих цифр. Таким образом, обрабатываемые данные с вероятностью 1 не соответствуют классической математико-статистической теории.

Как справиться с этим противоречием между теорией и практикой? В данной статье предлагается новый подход к анализу статистических данных с совпадениями.

Совпадения данных, модель их порождения и размножение выборок

В ранговых критериях анализируются не предъявленные для анализа результаты наблюдений, а их ранги. Если все результаты наблюдений различны, то их ранги различны, и мы находимся в классической ситуации, для которой разработаны различных ранговые критерии, исходящие из предположения о непрерывности функций распределения элементов выборок.

Если некоторые результаты измерений совпадают, образуют группу совпадающих значений, то результаты измерений в этой группе нельзя упорядочить, поскольку все они равны. Тем не менее желательно применить классические алгоритмы ранговой статистики. Чтобы выйти из противоречия между теорией и практикой, часто применяют эвристический приём – входят так называемые связанные ранги. Для этого

используют вариационный ряд, построенный по выборке, который представляет собой перестановку (упорядочение) элементов выборки в порядке неубывания. Для введения связанных рангов рассматривают номера (места) членов группы совпадающих значений в соответствующем вариационном ряду и приписывают каждому из этих значений один и тот же связанный ранг, равный среднему арифметическому номеров (мест), занимаемых членами этой группы в вариационном ряду.

Приведем пример.

Пример 1. Рассмотрим выборку, содержащую 26 элементов, записанных в два ряда для удобства дальнейших рассуждений:

$$17; 22; 3; 5; 15; 2; 0; 7; 13; 97; 66; 14; \quad (1)$$

$$47; 30; 2; 15; 1; 21; 25; 7; 44; 29; 33; 11; 6; 15. \quad (2)$$

Построенный по ней вариационный ряд имеет вид:

$$0 < 1 < 2 = 2 < 3 < 5 < 6 < 7 = 7 < 11 < 13 < 14 < 15 = 15 = \\ 15 < 17 < 21 < 22 < 25 < 29 < 30 < 33 < 44 < 47 < 66 < 97.$$

Для облегчения восприятия приведем эти данные в виде табл.1.

Таблица 1.

Номера в вариационном ряду, элементы выборки и ранги

Номера в вар. ряду	1	2	3	4	5	6	7	8	9	10	11	12	13
Ранги	1	2	3,5	3,5	5	6	7	8,5	8,5	10	11	12	14
Элементы выборки	0	1	2	2	3	5	6	7	7	11	13	14	15
Номера в вар. ряду	14	15	16	17	18	19	20	21	22	23	24	25	26
Ранги	14	14	16	17	18	19	20	21	22	23	24	25	26
Элементы	15	15	17	21	22	25	29	30	33	44	47	66	97

выборки													
---------	--	--	--	--	--	--	--	--	--	--	--	--	--

В выборке имеются три группы совпадающих значений. Результаты первой группы измерений (общее значение 2) встречаются дважды, они занимают в вариационном ряду выборки места 3 и 4, поэтому каждому из результатов измерений приписывается связанный ранг $(3+4)/2 = 3,5$. Во второй группе также из двух элементов (общее значение 7) занимаемые места – это 8 и 9, а потому каждому результату измерений в этой группе соответствует связанный ранг $(8+9)/2=8,5$. В третьей группе из трех элементов общее значение 15, результаты измерений из этой группы занимают места 13, 14 и 15, поэтому каждый из них имеет связанный ранг $(13+14+15)/3=14$. Для всех остальных элементов выборки ранг совпадает с номером (местом) в вариационном ряду. Приписанные элементам выборки ранги указаны в табл.1.

В алгоритмах расчётов значений ранговых критериев в качестве исходных данных используют ранги. Появление связанных рангов свидетельствует о том, что базовые предпосылки (непрерывность функций распределения) не выполнены. Распространенная практика использования связанных рангов не имеет теоретического обоснования, как и другие методы анализа данных с совпадениями (такие методы рассмотрены, например, в фундаментальной монографии [3]).

С целью снятия отмеченного противоречия между теоретическими результатами математической статистики и практикой анализа данных с совпадениями предлагаем ввести в рассмотрение **модели порождения данных**.

В математической статистике рассматривают выборки, элементы которых имеют непрерывные распределения. Таким образом, элемент выборки с вероятностью 1 является действительным числом с бесконечным числом значащих цифр. Подобные математические числа

нельзя использовать на практике. От них необходимо перейти к прагматическим числам с конечным числом значащих цифр. В рассмотренном выше примере в качестве элементов выборки рассмотрены неотрицательные целые числа с одной или двумя значащими цифрами.

Модель порождения данных должна описывать способ получения элементов выборки в виде математических чисел и переход от них к прагматическим числам.

Способ получения элементов выборки в виде математических чисел в ранговой статистике, как правило, состоит в том, что эти элементы являются независимыми одинаково распределенными случайными величинами, функция распределения которых непрерывна. Переход от таких чисел к прагматическим может осуществляться различными способами, рассмотренными в системной нечёткой интервальной математике [4]. Наиболее распространены способы, основанные на группировке [5]. Если в качестве прагматических чисел используются натуральные числа и 0, как в рассмотренном выше примере, то обычно полагают, что прагматическое число – это ближайшее целое число к математическому. Анализ процесса измерения может обосновать другие способы группировки (округления), а именно, прагматическое число – это целая часть математического числа, в другом варианте – целая часть плюс 1.

Рассмотрим подробнее округление к ближайшему целому для данных, приведённых в примере выше. Математическое число X переходит в прагматическое число N тогда и только тогда, когда $N - 0,5 < X < N + 0,5$ (вероятность попадания значения случайной величины X в счётное множество полуцелых значений X , т.е. таких, что $X = M + 0,5$ для некоторого целого числа M , равна 0, когда функция распределения непрерывна, поэтому полуцелыми значениями X можно пренебречь).

Для статистических данных, приведённых в табл.1, на местах 3 и 4 могут быть, например, значения 1,9 и 2,2, переходящие после округления в общее значение 2. На местах 8 и 9 – результаты измерений 6,8 и 7,3, переходящие после округления в общее значение 7. На местах 8 и 9 – результаты измерений 6,8 и 7,3, переходящие после округления в общее значение 7. На местах 13, 14 и 15 могут быть, например, значения 14,9, 15,1 и 15,4, переходящие после округления в общее значение 15. Таким образом, исчезают группы связанных рангов, происходит «размножение выборок».

Двухвыборочный критерий Вилкоксона

Влияние «размножения выборок» на свойства ранговых критериев проследим на примере двухвыборочного критерия Вилкоксона. Поскольку в литературе имеются различные понимания термина «критерий Вилкоксона», то начнём с описания этого критерия.

В прикладных исследованиях часто возникает необходимость выяснить, различаются ли генеральные совокупности, из которых взяты две независимые выборки. Например, в технических исследованиях – надо выяснить, влияет ли способ упаковки подшипников на их потребительские качества через год после хранения. Результат проверки используется при выборе способа упаковки рассматриваемого типа подшипников. В маркетинге – есть ли отличия в потребительском поведении мужчин и женщин. Если отличается – рекламные ролики и плакаты надо делать отдельно для мужчин и отдельно для женщин. Если нет – рекламная кампания может быть единой.

В математико-статистических терминах постановка задачи такова: имеются две выборки $x_1, x_2, \dots, x_m$ и $y_1, y_2, \dots, y_n$ (т. е. наборы из m и n действительных чисел), требуется проверить их однородность. Термин «однородность» уточняется ниже.

Противоположным понятием является «различие» (другими словами, «наличие эффекта при переходе от одной выборки к другой»). С прикладной точки зрения полезно изменить постановку задачи: требуется проверить, есть ли различие между выборками. Если различия нет, то для дальнейшего изучения две рассматриваемые выборки часто объединяют в одну.

Например, в маркетинге важно выделить сегменты потребительского рынка. Если обоснована однородность двух выборок, то целесообразно объединение сегментов, из которых они взяты, в один. В дальнейшем это позволит осуществлять по отношению к ним одинаковую маркетинговую политику (проводить одни и те же рекламные мероприятия и т.п.). Если же установлено различие, то поведение потребителей в двух сегментах существенно различается, а потому объединять эти сегменты в один нельзя, для каждого из них может понадобиться своя маркетинговая стратегия.

В прикладной математической статистике обычно рассматривают следующую вероятностную модель двух независимых выборок числовых результатов наблюдений. Первая выборка описывается набором m случайных величин $X_1, X_2, \dots, X_m$, имеющих одну и ту же функцию распределения $F(x)$, а вторая выборка - набором n случайных величин $Y_1, Y_2, \dots, Y_n$, имеющих одну и ту же функцию распределения $G(x)$, причем все эти $m + n$ случайных величин $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ независимы в совокупности. Без ограничения общности можно считать, что $m \leq n$, в противном случае выборки достаточно поменять местами. Как правило, предполагается, что функции $F(x)$ и $G(x)$ непрерывны и строго возрастают. Из непрерывности этих функций следует, что с вероятностью 1 все $m + n$ результатов наблюдений различны. В реальных статистических данных иногда встречаются совпадения, но сам факт их наличия - свидетельство

нарушений предпосылок только что описанной базовой математической модели.

Предполагается, что результаты наблюдений (измерений, испытаний, анализов, опытов) – числа $x_1, x_2, \dots, x_m$ и $y_1, y_2, \dots, y_n$ – являются реализациями случайных величин $X_1, X_2, \dots, X_m$ и $Y_1, Y_2, \dots, Y_n$ соответственно, т.е. значениями случайных величин (как функций от элементарного события ω), принимаемыми для определенного случайного события ω .

В прикладных исследованиях, примеры которых перечислены выше, под «однородностью» понимают тождественное равенство функций распределения $F(x)$ и $G(x)$. В этом и состоит нулевая гипотеза, для проверки которой разработан ряд статистических критериев [6, гл.2]. Используют также термин «абсолютная однородность» с целью разграничить задачи проверки статистических гипотез абсолютной однородности и однородности по параметрам, т.е. равенства математических ожиданий, дисперсий или иных характеристик, например, коэффициентов вариации для двух выборок.

Статистика S двухвыборочного критерия Вилкоксона для проверки однородности двух независимых выборок рассчитывается следующим образом. Выборки объединяют вместе, получают т.н. «объединённую выборку». Все элементы объединенной выборки $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ упорядочивают в порядке неубывания, строя общий (объединенный) вариационный ряд. Элементы первой выборки $X_1, X_2, \dots, X_m$ занимают в общем вариационном ряду места с номерами $R_1, R_2, \dots, R_m$, другими словами, имеют ранги $R_1, R_2, \dots, R_m$. Тогда статистика Вилкоксона — это сумма рангов элементов первой выборки

$$S = R_1 + R_2 + \dots + R_m.$$

Эта статистика была предложена и изучена Вилкоксоном в 1945 г. [7].

Вскоре (в 1947 г.) Манн и Уитни разработали метод проверки однородности независимых выборок на основе статистики, которая носит теперь их имена [8]. Статистика U Манна - Уитни определяется как число пар (X_i, Y_j) таких, что $X_i < Y_j$, среди всех mn пар, в которых первый элемент - из первой выборки, а второй - из второй. Как известно (см., например, [3, с.160]),

$$U = mn + m(m+1)/2 - S.$$

Поскольку S и U линейно связаны, то в литературе часто говорят не о двух критериях – Вилкоксона и Манна - Уитни, а об одном - критерии Вилкоксона (Манна-Уитни) или Вилкоксона – Манна - Уитни. Иногда критерий Вилкоксона называют критерием Манна – Уитни. Мы используем терминологию, закреплённую в фундаментальном справочнике Л.Н. Большева и Н.В. Смирнова [9].

Критерий Вилкоксона - один из самых известных инструментов непараметрической статистики (наряду с критериями на основе статистик типа Колмогорова - Смирнова, омега-квадрат [10] и коэффициентами ранговой корреляции). Свойствам этого критерия и таблицам его критических значений уделяется место во многих монографиях по математической и прикладной статистике (см., например, [3, 9, 11]). Отметим, что в литературе имеются и неточные утверждения (мифы) относительно возможностей критерия Вилкоксона. Два таких мифа проанализированы в статье [12].

Если выборки абсолютно однородны, т.е. их функции распределения $F(x)$ и $G(x)$ тождественно совпадают, то справедлива нулевая гипотеза

$$H_0: F(x) = G(x) \text{ при всех } x. \quad (3)$$

Эту гипотезу можно проверять, сравнивая значение статистики Вилкоксона S с граничными значениями, приведёнными в соответствующих таблицах (см., например, [9, табл. 6.8]), включённых к настоящему времени в программные продукты, предназначенные для

статистического анализа данных. Однако в них ограничен диапазон возможных значений объемов выборок. Альтернатива – воспользоваться асимптотической нормальностью статистики Вилкоксона S .

Целесообразно строить алгоритм принятия решения на основе центрированной нормированной статистики Вилкоксона

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

Согласно [3, 9] математическое ожидание $M(S)$ статистики S равно

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

а дисперсия $M(S)$ статистики S равна

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

В [3] доказана асимптотическая нормальность статистики T : при безграничном росте объемов выборок ($m \rightarrow \infty$, $n \rightarrow \infty$) функция распределения $P(T < x)$ статистики T стремится к функции $\Phi(x)$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Асимптотическая нормальность статистики T позволяет обычным образом [13] задавать критические значения критерия проверки однородности двух независимых выборок, построенного на основе этой статистики, с помощью квантилей стандартного нормального распределения.

Пример 2. Рассмотрим данные, приведённые в примере 1, а именно, первую выборку (1) объёма $m = 12$ и вторую выборку (2) объёма $n = 14$. Информация об объединённой выборке объёма $m + n$ приведена в табл. 1. Для расчёта значения статистики Вилкоксона не нужны численные значения элементов объединённой выборки, а также их номера в вариационном ряду. Однако нужна дополнительная информация – к какой выборке (первой или второй) относится элемент с определённым рангом. Сведём необходимые данные в табл. 2.

Таблица 2.

Ранги и номера выборок для данных таблицы 1

Ранги	1	2	3,5	3,5	5	6	7	8,5	8,5	10	11	12	14
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
Номера выборок	2	2	1	2	1	2	2	2	2	2	2	1	1

Найдём значение статистики Вилкоксона, т.е. сумму рангов элементов первой выборки

$$S = R_1 + R_2 + \dots + R_m = 1+3,5+5+6+8,5+11+12+14+16+18+25+26=146.$$

Вычислим 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. Метод контроля основан на 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, необходимо, чтобы

$$S + S_1 = (m + n)(m + n + 1)/2 = (12+14)(12+14+1)/2 = 351.$$

В соответствии с ранее проведенными расчетами $S + S_1 = 146 + 205 = 351$.
Необходимое условие правильности расчетов выполнено.

В соответствии со сказанным ранее

$$M(S) = 12(12+14+1)/2 = 162, D(S) = 12 \cdot 14(12+14+1)/12 = 378.$$

Следовательно,

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

Поскольку $|T| \leq 1,96$, то гипотеза однородности принимается на уровне значимости 0,05.

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

$$M(S_1) = 14(12+14+1)/2 = 189, D(S) = D(S_1) = 378,$$

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

Таким образом, значения статистики критерия проверки однородности отличаются только знаком (легко 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. Поэтому для уменьшения объема таблиц критических значений ранее было принято считать первой выборку меньшего объема. К настоящему времени в связи с развитием программных средств анализа статистических данных эта рекомендация утратила актуальность.

Анализ данных с совпадениями при использовании критерия Вилкоксона

Продemonстрируем процедуру анализа данных с совпадениями на примере численных значений элементов двух независимых выборок, рассмотренных в примерах 1 и 2 (см. табл. 1 и 2 выше). Построим «облако» упорядочений (ранжировок), соответствующее введенной выше модели порождения данных при анализе совпадающих значений. Поскольку в выборки входят целые числа, то, как отмечалось выше, достаточно принять, что максимальная абсолютная погрешность не превосходит $\Delta < 0,5$. Результаты построения «облака» упорядочений

приведены в табл. 3. В ней рассмотрены три группы совпадающих значений, для которых ранее в табл. 1 были введены связанные ранги.

Значение 2 имеется как в выборке 1, так и в выборке 2. При построении «облака» меньшим среди реальных значений, округлённых до числа 2, может быть либо реальное значение из выборки 1 (упорядочения 1 - 6), либо реальное значение из выборки 2 (упорядочения 7 - 12), всего 2 возможности.

Аналогична ситуация со значением 7. Оно имеется как в выборке 1, так и в выборке 2. При построении «облака» меньшим может быть либо реальное значение из выборки 1 (упорядочения 1 - 3 и 7 - 9 в табл. 3), либо реальное значение из выборки 2 (упорядочения 4 - 6 и 10 - 12), всего 2 возможности.

Несколько сложнее со значением 15. Оно один раз встречается в выборке 1 и два раза в выборке 2. Следовательно, необходимо рассмотреть упорядочение трех реальных значений. При построении «облака» реальное значение из выборки 1 может стоять либо на первом месте (упорядочения 1, 4, 7, 10), либо на втором месте (упорядочения 2, 5, 8, 11), либо на третьем (упорядочения 3, 6, 9, 12), остальные два места занимают реальные значения их второй выборки, всего 3 возможности для рассматриваемой группы совпадающих значений.

Сводя вместе результаты анализа трех групп совпадающих значений, заключаем, что «облако» состоит из $2 \times 2 \times 3 = 12$ упорядочений. Они приведены в табл. 3.

Таблица 3.

«Облако» упорядочений для совпадающих данных табл. 1 и 2.

Связанные ранги	3,5	3,5	8,5	8,5	14	14	14	Статистика S
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	

Подсчитаем для упорядочений в табл.3 значения статистики Вилкоксона, т.е. сумму рангов элементов первой выборки. Как показано выше, при использовании связанных 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 элементы объединенной выборки независимы в совокупности, то все 12 упорядочений равновероятны, и распределение статистики Вилкоксона S имеет вид, представленный в табл. 4. Значение этой статистики, соответствующее использованию связанных рангов, т.е. число 146, является центральным.

Таблица 4.

Распределение значений статистики Вилкоксона в «облаке»

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, мало влияние на выводы согласно теории проверки статистических гипотез. Действительно, значение центрированной и нормированной статистики Вилкоксона T лежит в интервале $[T_1, T_2]$, где

$$T_1 = (144 - 162) / 19,44 = -0,93, T_2 = (148 - 162) / 19,44 = -0,72.$$

Разброс небольшой, вывод о принятии нулевой гипотезы об однородности сохраняется.

Дополнительный анализ требуется в случае, когда минимальное из возможных значений статистики Вилкоксона T_1 лежит ниже граничного значения, а максимальное T_2 – выше. Отметим, что это замечание аналогично заключению о выгодности инвестиционного проекта на основе использования значения чистой текущей стоимости NPV с учётом погрешности этого значения [14].

Критерий знаков для проверки однородности связанных выборок

Модель порождения данных при анализе связанных выборок такова. Статистические данные имеют вид (x_i, y_i) , $i = 1, 2, \dots, n$, где x_i – результат измерения первого параметра, y_i – второго, n – объем выборки (число пар измерений). В примере, рассмотренном в работе [15], x_i – результат измерения вязкости на нестандартном приборе фабрики им. Петрова в i -ой партии, а y_i – результат измерения вязкости на стандартном вискозиметре ВЗ-4 в той же i -ой партии. Пусть a_i – истинное значение показателя вязкости в i -ой партии. Естественно считать, что указанные выше случайные вектора независимы в совокупности. При этом они не являются

одинаково распределенными, поскольку отличаются истинными значениями показателей качества a_i . Принимаем, что *при каждом i случайные величины $x_i - a_i$ и $y_i - a_i$ независимы и одинаково распределены.* Это условие и означает *однородность связанных выборок*. Параметры связи - величины a_i . Их наличие не позволяет объединить первые координаты в одну выборку, вторую - во вторую, как делается в случае проверки однородности двух независимых выборок [1, 2].

В предположении непрерывности функций распределения из условия однородности в связанных выборках вытекает, что

$$P(x_i < y_i) = P(x_i \geq y_i) = \frac{1}{2}. \quad (4)$$

Рассмотрим случайные величины $Z_i = x_i - y_i$, $i = 1, 2, \dots, n$. Из соотношения (4) вытекает, что при справедливости гипотезы однородности для связанных выборок эти случайные величины имеют нулевые медианы. Другими словами, в примере, рассмотренном в работе [15], вместо проверки того, что методы измерения вязкости дают схожие результаты, можно проверять равенство 0 медиан величин Z_i , т.е. проверять следствие (4) гипотезы однородности связанных выборок.

Для проверки гипотезы о том, что медианы величин Z_i нулевые, можно применить широко известный критерий знаков (см., например, справочник [2, с.89-91]). Согласно этому критерию необходимо подсчитать, в скольких партиях $x_i < y_i$ и в скольких $x_i \geq y_i$. В [15] для представленных химическим комбинатом данных $x_i < y_i$ в 187 случаях из 213 и $x_i \geq y_i$ в 26 случаях из $n = 213$.

Если рассматриваемая гипотеза однородности верна, то число W осуществлений события $x_i < y_i$, т.е. $Z_i < 0$, имеет биномиальное распределение с параметрами $p = 1/2$ и n . Для данных химического комбината $n = 213$, математическое ожидание $M(W)=106,5$, а среднее квадратическое отклонение $\sigma = \sqrt{np(1-p)} = 7,3$. Следовательно, интервал

$M(W) \pm 3\sigma$ - это интервал $84 \leq W \leq 129$. Найденное по данным химического комбината значение $W=187$ лежит далеко вне этого интервала. Поэтому рассматриваемую гипотезу необходимо отвергнуть (на любом используемом в прикладных работах уровне значимости, в частности, на уровне значимости 1%).

Наиболее простая модель проверки однородности для связанных выборок сводится к проверке гипотезы о значении параметра биномиального распределения:

$$H_0 : p = \frac{1}{2},$$

где $p = P\{x_i < y_i\}$.

В этой модели применяют «критерий знаков» [1, 2, 15], основанный на расчёте числа W осуществлений события $\{x_i < y_i\}$. При справедливости гипотезы однородности эта случайная величина имеет биномиальное распределение с вероятностью успеха $p = 1/2$ и числом испытаний n . Альтернативная гипотеза состоит в том, что вероятность успеха отличается от $1/2$:

$$H_1 : p \neq \frac{1}{2}.$$

Гипотезу $p = 1/2$ можно проверять как непосредственно с помощью биномиального распределения (используя таблицы или программное обеспечение), так и опираясь на (интегральную) теорему Муавра-Лапласа. Согласно этой теореме

$$\lim_{n \rightarrow \infty} P\left\{\frac{2W - n}{\sqrt{n}} \leq x\right\} = \Phi(x)$$

при всех x , где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из теоремы Муавра-Лапласа вытекает правило принятия решений на уровне значимости 5%: если

$$\left| \frac{2W - n}{\sqrt{n}} \right| \leq 1,96,$$

то гипотезу однородности связанных выборок принимают, в противном случае отклоняют.

Замечание. Здесь и далее при желании использовать другой уровень значимости применяют в качестве критического значения иной квантиль нормального распределения. Использование предельных теорем допустимо при достаточно больших объемах выборки. По поводу придания точного смысла термину «достаточно большой» в науке продолжаются дискуссии. Обычно считается, что несколько десятков (два-три десятка) — это уже «достаточно много». Более правильно констатировать, что ответ зависит от задачи, от ее сложности и практической значимости.

Как и в случае двухвыборочного критерия Вилкоксона, при применении критерия знаков проявляется эффект влияния совпадений округлённых значений. А именно, эффект, вызванный появлением значения 0 среди $Z_i = x_i - y_i$, $i = 1, 2, \dots, n$.

В соответствии с теорией математической статистики вероятность осуществления равенства $x_i - y_i$ равна 0, поскольку независимые случайные величины x_i и y_i имеют непрерывные функции распределения. Однако в практике статистического анализа реальных данных достаточно часто встречаются значения $Z_i = 0$. Это происходит из-за округления.

Как влияет округление на принятие решений на основе критерия знаков? Поскольку решение принимается на основе числа W осуществлений события $\{x_i < y_i\}$, то в случае, если до округления была случайная величина $x_i - y_i$, округлённая затем до 0, принимала отрицательное значение, то значение статистики критерия W должно быть увеличено на 1. Если же величина $x_i - y_i$ была положительной, то значение статистики критерия W не меняется. Следовательно, если округлённое

значение $Z_i = 0$ встречается k раз, то рассчитанное по данным до округления число осуществлений событий $\{x_i < y_i\}$ может принимать любое натуральное число от W до $W + k$ (здесь W – результат подсчёта до округления). Таким образом, для обоснованного принятия решения надо рассматривать два значения статистики критерия - W и $W + k$. Вместе с тем ясно, что различие между выводами на основе этих двух значений имеется лишь тогда, когда W близко к границе критической области.

Иногда используют модификацию критерия знаков на основе подсчёта числа W_1 осуществлений события $\{x_i \leq y_i\}$. Ясно, что $W_1 = W + k$, где k – число округлённых значений $Z_i = 0$. Из сказанного выше ясно, что при использовании рассматриваемой модификации критерия знаков для обоснованного принятия решения надо рассматривать два значения статистики критерия, а именно, W_1 и $W_1 - k$.

Заключение

Данная статья посвящена анализу особенностей применения ранговых критериев непараметрической статистики в условиях несоблюдения важного теоретического предположения о том, что функции распределения результатов наблюдений непрерывны. Отмечаем, что для реальных выборок такое предположение редко выполняется и достаточно часто имеются совпадающие значения. Это заставляет ряд исследователей вводить специальные поправки, учитывающие наличие «связанных рангов». В данной статье предложен другой подход – на основе теории устойчивости статистических процедур [16] рассчитать интервал возможных значений ранговой статистики. Для пояснения подхода рассматривается базовый пример, в котором проводится проверка однородности функций распределения двух независимых выборок с помощью двухвыборочного критерия Вилкоксона (Манна - Уитни).

Вместо обычно применяемой точечной оценки в данной работе получен интервал значений статистики Вилкоксона.

Предлагается следующий алгоритм анализа совпадений при расчете непараметрических ранговых статистик (см. также [17]).

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 $y_1 = y_2 = \dots = y_k = a$ получена (при связанным с процессом измерения округлениях) из набора $x_j = y_j + \varepsilon_j$, $j = 1, 2, \dots, k$, при некоторых малых отклонениях ε_j , $j = 1, 2, \dots, k$, среди которых нет совпадающих. Одна точка a соответствует «облаку» из k различных точек $x_j = a + \varepsilon_j$, $j = 1, 2, \dots, k$.

3. Упорядочить элементы «облака».

Комментарий к п. 3 алгоритма. Для группы совпадающих значений $y_1 = y_2 = \dots = y_k = a$ 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 (см. п.3 алгоритма выше). Итогом выполнения пп. 1- 4 является «размножение» единой исходной объединенной выборки на некоторое число похожих, отличающихся введением того или иного

упорядочения совпадающих значений (см. выше п.2 алгоритма). Другими словами, основная идея алгоритма данной работы лежит в общем русле использования различных процедур «размножения выборки» (бутстреп, метод складного ножа и др.) [1].

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

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

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 значений» в соответствии с вероятностями их появления. Нам представляется более практичным подход на основе системной нечёткой интервальной математики [4] и теории устойчивости статистических процедур [15], в простейшем случае сводящийся к расчету размаха возможных значений ранговой статистики (т.е. значений, входящих в «облако»).

Второй пример – анализ влияния совпадений при применении критерия знаков для проверки однородности связанных выборок.

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

Литература

1. Орлов А.И. Прикладная статистика. - М.: Экзамен, 2006. - 671 с.
2. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. Изд. 3-е. – М.: Наука, 1983. - 416 с.
3. Гаек Я., Шидак З. Теория ранговых критериев. - М.: Наука, 1971. - 376 с.
4. Орлов А.И., Луценко Е.В. Системная нечеткая интервальная математика. Монография (научное издание). – Краснодар, КубГАУ. 2014. – 600 с.
5. Орлов А.И. Статистическое оценивание для сгруппированных данных // Научный журнал КубГАУ. 2014. №98. С. 1097–1117.
6. Орлов А.И. Эконометрика. Изд. 4-е, доп. и перераб. Учебник для вузов. — Ростов-на-Дону: Феникс, 2009. — 572 с.
7. Wilcoxon F. Individual comparisons by ranking methods // Biometrics. 1945. N 1. P. 80-83.
8. Mann H.B., Whitney D.R. On test whether one of the two random variables is stochastically larger than the other // Annals of Mathematical Statistics. 1947. V. 18. P. 50-60.
9. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. Изд. 3-е.— М.: Наука, 1983. – 416 с.
10. Орлов А.И. Непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат и ошибки при их применении // Научный журнал КубГАУ. 2014. №97. С. 647–675.
11. Холлендер М., Вульф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. – 518 с.
12. Орлов А.И. Двухвыборочный критерий Вилкоксона – анализ двух мифов // Научный журнал КубГАУ. 2014. № 104. С. 91 – 111.
13. Орлов А.И. Вероятность и прикладная статистика: основные факты: справочник. — М.: КноРус, 2009. — 192 с.
14. Орлов А.И. Оценка погрешностей характеристик финансовых потоков инвестиционных проектов в ракетно-космической промышленности // Научный журнал КубГАУ 2015. №109. С. 238 – 264.
15. Орлов А.И. О проверке однородности связанных выборок // Научный журнал КубГАУ. 2016. №123. С. 708 – 726.
16. Орлов А.И. Устойчивые экономико-математические методы и модели: монография. — М.: Ай Пи Ар Медиа, 2022. — 337 с.
17. Орлов А.И. Модель анализа совпадений при расчете непараметрических ранговых статистик // Заводская лаборатория. Диагностика материалов. 2017. Т.83. №11. С. 66-72.

References

1. Orlov A.I. Prikladnaya statistika. - M.: Ekzamen, 2006. - 671 s.
2. Bol'shev L.N., Smirnov N.V. Tablicy matematicheskoy statistiki. Izd. 3-e. – M.: Nauka, 1983. - 416 s.
3. Gaek YA., SHidak Z. Teoriya rangovykh kriteriev. - M.: Nauka, 1971. - 376 s.
4. Orlov A.I., Lucenko E.V. Sistemnaya nechetkaya interval'naya matematika. Monografiya (nauchnoe izdanie). – Krasnodar, KubGAU. 2014. – 600 s.
5. Orlov A.I. Statisticheskoe ocenivanie dlya sgruppированных данных // Nauchnyj zhurnal KubGAU. 2014. №98. S. 1097–1117.
6. Orlov A.I. Ekonometrika. Izd. 4-e, dop. i pererab. Uchebnik dlya vuzov. — Rostov-na-Donu: Feniks, 2009. — 572 s.
7. Wilcoxon F. Individual comparisons by ranking methods // Biometrics. 1945. N 1. P. 80-83.
8. Mann H.B., Whitney D.R. On test whether one of the two random variables is stochastically larger than the other // Annals of Mathematical Statistics. 1947. V. 18. P. 50-60.
9. Bol'shev L.N., Smirnov N.V. Tablicy matematicheskoy statistiki. Izd. 3-e.– M.: Nauka, 1983. – 416 s.
10. Orlov A.I. Neparametricheskie kriterii soglasiya Kolmogorova, Smirnova, omega-kvadrat i oshibki pri ih primenenii // Nauchnyj zhurnal KubGAU. 2014. №97. S. 647–675.
11. Hollender M., Vul'f D. Neparametricheskie metody statistiki. – M.: Finansy i statistika, 1983. – 518 s.
12. Orlov A.I. Dvuhvyborochnyj kriterij Vilkoksona – analiz dvuh mifov // Nauchnyj zhurnal KubGAU. 2014. № 104. S. 91 – 111.
13. Orlov A.I. Veroyatnost' i prikladnaya statistika: osnovnye fakty: spravochnik. — M.: KnoRus, 2009. — 192 s.
14. Orlov A.I. Ocenka pogreshnostej harakteristik finansovykh potokov investicionnykh proektov v raketno-kosmicheskoy promyshlennosti // Nauchnyj zhurnal KubGAU 2015. №109. S. 238 – 264.
15. Orlov A.I. O proverke odnorodnosti svyazannykh vyborok // Nauchnyj zhurnal KubGAU. 2016. №123. S. 708 – 726.
16. Orlov A.I. Ustojchivye ekonomiko-matematicheskie metody i modeli: monografiya. — M.: Aj Pi Ar Media, 2022. — 337 c.
17. Orlov A.I. Model' analiza sovpadenij pri raschete neparametricheskikh rangovykh statistik // Zavodskaya laboratoriya. Diagnostika materialov. 2017. T.83. №11. S. 66-72.