Математические методы исследования
Mathematical methods of investigation

DOI: https://doi.org/10.26896/1028-6861-2023-89-11-98-106

О ТРЕБОВАНИЯХ К СТАТИСТИЧЕСКИМ МЕТОДАМ АНАЛИЗА ДАННЫХ (ОБОБЩАЮЩАЯ СТАТЬЯ)

© **Александр Иванович Орлов**

Московский государственный технический университет им. Н. Э. Баумана, Россия, 105005, Москва, Бауманская 2-я, д. 5; e-mail: prof-orlov@mail.ru

Статья поступила 29 декабря 2022 г. Поступила после доработки 27 января 2023 г. Принята к публикации 28 февраля 2023 г.

Консультируя в течение полувека научных работников различных специальностей, рецензируя их статьи и книги, оппонируя диссертации, автор познакомился с сотнями конкретных исследований по разработке и применению статистических методов. В итоге выявил разнообразные недостатки при проведении исследований и в публикациях их результатов, которые мешают восприятию смысла, а в ряде случаев ставят под сомнение адекватность выводов. Поэтому целесообразно выработать естественные требования к методам обработки данных и представлению результатов статистического анализа данных. Данная работа посвящена первоначальному рассмотрению ряда формулировок таких требований. Исходим из современной парадигмы прикладной статистики, основанной на непараметрической и нечисловой статистике и сменившей примитивную парадигму XIX в. и парадигму середины XX в., использующую параметрические системы распределений. При описании и обсуждении процедур анализа статистических данных начинать надо с вероятностно-статистических моделей порождения изучаемых данных. Анализ многообразия моделей регрессионного анализа приводит к выводу, что не существует единой «стандартной модели». Необходимо исходить из теории измерений, согласно которой первый шаг при анализе данных — выявление шкал, в которых они измерены. Статистические выводы должны быть инвариантны относительно допустимых преобразований шкал измерения данных. Поскольку практически все распределения реальных данных ненормальны, предпочтения следует отдавать непараметрическим постановкам. Возможность применения параметрических семейств распределений должна быть тщательно обоснована. В соответствии с теорией проверки статистических гипотез должны быть указаны как нулевая, так и альтернативная гипотезы. Необходимо изучение устойчивости выводов, получаемых на основе модели, относительно допустимых изменений исходных данных и предпосылок модели. Проблемам разработки системы требований к статистическим моделям и методам будет посвящен ряд дальнейших публикаций автора.

Ключевые слова: математические методы исследования; статистические методы; анализ данных; вероятностно-статистическая модель; прикладная статистика; непараметрика; нечисловая статистика; теория измерений; регрессионный анализ.

ON THE REQUIREMENTS FOR STATISTICAL METHODS OF DATA ANALYSIS (GENERALIZING ARTICLE)

© **Alexander I. Orlov**

Bauman Moscow State Technical University, 5, 2-ya Baumanskaya ul., Moscow, 105005, Russia; e-mail: prof-orlov@mail.ru

Received December 29, 2022. Revised January 27, 2023. Accepted February 28, 2023.

Half a century activity in consulting researchers of various specialties, reviewing articles and books, opposing dissertations provided the possibility of getting acquainted with hundreds of specific studies on the development and application of statistical methods. A variety of shortcomings in conducting research and publishing the results of studies have been revealed, which hinder the perception of the data obtained, and in some cases cast doubt on the adequacy of the conclusions. Therefore, it appeared advisable to develop natural requirements for methods of data processing and presentation of the results of statistical data analysis. This study is devoted to an initial consideration of a number of formulations of such require-

ments. We proceed from the modern paradigm of applied statistics, based on non-parametric and non-numerical statistics which replace the primitive paradigm of the 19th century and the paradigm of the middle of the 20th century using parametric distribution systems. When describing and discussing the procedures for analyzing statistical data, it is necessary to start with probabilistic-statistical models for generating the data under study. It is necessary to proceed from the theory of measurements, according to which the first step in the data analysis is identification of the scales in which they are measured. Statistical inference must be invariant under the allowable transformations of data measurement scales. Since almost all distributions of real data are non-normal, a preference should be given to non-parametric formulations. The possibility of using parametric families of distributions must be carefully justified. In accordance with the theory of testing statistical hypothesis, both the null and alternative hypotheses must be specified. It is necessary to study the stability of the conclusions drawn from the model with respect to acceptable changes in the initial data and assumptions of the model. A number of further publications of the author will be devoted to the problems of developing a system of requirements for statistical models and methods.

Keywords: mathematical research methods; statistical methods; data analysis; probabilistic-statistical model; applied statistics; non-parametric statistics; non-numerical statistics; measurement theory; regression analysis.

Введение

Консультируя в течение полувека научных работников различных специальностей, рецензируя их статьи и книги, оппонируя диссертации, автор познакомился с сотнями конкретных исследований по разработке и применению статистических методов. Критический анализ накопленного материала позволил разработать общий подход к проведению таких исследований и ряд частных методов [1]. Выявлены разнообразные недостатки при проведении исследований и в публикациях их результатов, которые мешают восприятию смысла, а в ряде случаев ставят под сомнение адекватность выводов. Поэтому целесообразно сформулировать и обсудить естественные требования к методам обработки данных и представлению результатов статистического анализа конкретных данных. В статье [2] сделана попытка выделить основные характеристики методов прикладной статистики и сформулировать требования к этим методам (т.е. к значениям упомянутых характеристик методов). Например, одно из требований: статистические выводы должны быть инвариантны относительно допустимых преобразований шкал измерения.

В целях стандартизации математических орудий (пользуемся терминологией Н. Бурбаки [3, с. 253]) представляется целесообразным развернуть работу по сертификации статистических методов и соответствующих пакетов программ, а также учебных курсов и материалов. Однако стандартизация полезна только тогда, когда она проводится квалифицированными специалистами, в противном случае вместо пользы получаем вред. Примером является печальная судьба многообразия стандартов по статистическим методам управления качеством, большую часть которых пришлось отменить из-за ошибок разработчиков.

Настоящая работа посвящена первоначальному рассмотрению ряда формулировок требований к методам обработки данных и представлению результатов статистического анализа конкретных данных.

Постановка проблемы

Отечественная научная школа в области статистических методов анализа данных опирается на аксиоматику теории вероятностей А. Н. Колмогорова [4]. В послевоенные годы исследователи исходили из монографии Г. Крамера [5]. В состав Академии наук СССР входили только два специалиста по математической статистике — члены-корреспонденты Н. В. Смирнов и Л. Н. Большев. Они выпустили монографию «Таблицы математической статистики» [6], которая также содержала развернутые пояснения. Эта монография стала одним из высших достижений отечественной статистики XX в. Другое достижение — фундаментальная энциклопедия «Вероятность и математическая статистика» [7]. Для будущих математиков предназначены учебники акад. А. А. Боровкова [8, 9], задачник [10], для студентов высших технических учебных заведений — учебные пособия [11, 12]. Были выпущены сотни добротных отечественных изданий и переводов.

На этой базе могли быть подготовлены учебные издания по статистическим методам анализа данных, соответствующие современному научному уровню. К сожалению, этого не произошло — вывод сделан на основе анализа учебных изданий, информация о которых появляется при поисковом запросе «Теория вероятностей и математическая статистика» [13 – 18]. Приведем некоторые обнаруженные недостатки и сопоставим их с соответствующими разделами учебника [1] (первое издание вышло в 2006 г.).

Исследователи часто игнорируют необходимость формирования и обоснования вероятностно-статистических моделей порождения данных [1, разд. 5.1]. В частности, не видят [13, 14] прин-

ципиальной разницы между двумя моделями выборки — представительной выборкой из конечной совокупности и последовательностью независимых одинаково распределенных случайных величин [1, разд. 2.4]. Не понимают [16] разницы между регрессионной моделью с детерминированной независимой переменной и моделью, в которой исходные данные — выборка из двумерного нормального распределения [1, гл. 9]. Метод наименьших квадратов рассматривают на рецептурном уровне, без вывода формул для оценок параметров [14] и без доверительных оценок для зависимости [13]. О существовании непараметрической регрессии не подозревают [14].

Многие считают аксиомой нормальное распределение результатов измерений. В основополагающей книге В. В. Налимова [19], выпущенной в 1960 г., разъяснено, что распределения реальных данных, как правило, не являются нормальными. Более подробно это фундаментальное положение обсуждается в [1, разд. 5.1]. Однако в [16 – 18] принимают нормальность данных без какого-либо обоснования. Авторы [14] игнорируют тот факт, что применение критерия Стьюдента основано на предположении нормальности распределения элементов выборки, и тем более не пытаются проверить это предположение. Согласно [14] «предположение о нормальности основывается на центральной предельной теореме, в соответствии с которой случайные величины, являющиеся суммой большого числа других случайных факторов (здесь пропущено условие независимости факторов — А. О.), ни один из которых не является доминирующим, имеют приближенно нормальное распределение». Авторы [14] не подозревают, что если факторы действуют не аддитивно, а мультипликативно, то согласно той же центральной предельной теореме результирующая величина имеет логарифмически нормальное распределение, а не нормальное [1, разд. 5.1].

В этой же работе [14] разбирают построение безнадежно устаревших гистограмм, хотя уже с середины XX в. для оценивания плотности используют непараметрические ядерные оценки [1, разд. 5.6]. Авторы не замечают, что переход от вариационного ряда к гистограммам (группировка) влечет потерю информации [16]. Приверженность к гистограммам приводит к тому, что проверку согласия опытного распределения с теоретическим проводят с помощью критерия хи-квадрат [14, 15, 17], т.е. проверяют более слабую гипотезу о вероятностях попадания в интервалы группирования.

В анализируемых методических материалах много сравнительно мелких ошибок. Так, в них даны неправильные определения выборочной медианы [14, с. 67], эмпирической функции распределения ([14, с. 73], [16, с. 22]), интервальной оценки [14, с. 79], параметрической и непараметрической гипотез [16, с. 28]. Авторы [14] не знают, как выглядит график плотности нормального распределения, кроме того, они ошибочно полагают, что из равенства нулю коэффициента корреляции между двумя случайными величинами следует их независимость [14, с. 105]. Профессионал пишет «теория вероятностей», профан — «теория вероятности» [16, с. 39].

Есть сомнительные места и у более квалифицированных авторов. Так, для оценки параметров распределения рекомендуют метод максимального правдоподобия [15, 18], хотя с современной точки зрения следует применять метод одношаговых оценок [1, разд. 6.2]. В [13] приведен доверительный интервал для математического ожидания, но не указано, что он является асимптотическим, отсутствует ссылка на центральную предельную теорему. Не разъяснено, в чем специфика использования квантилей нормального закона и распределения Стьюдента. Авторы [13] не ведают, какие критические значения следует использовать [1, разд. 2.6] при применении критерия Колмогорова для проверки согласия с нормальным семейством распределения (и тем более не знают, например, про критерий Шапиро – Уилка для проверки нормальности). Им неизвестно, как строить доверительный интервал для коэффициента корреляции в общем случае [1, разд. 9.1], предложенный еще Г. Крамером [5]. Они не знают, что для проверки равенства математических ожиданий по двум независимым выборкам надо использовать не критерий Стьюдента, а критерий Крамера – Уэлча [1, разд. 8.2]. Вопреки [13] с помощью критерия Вилкоксона нельзя проверить гипотезу о совпадении функций распределения [1, разд. 8.3].

В настоящее время для описания неопределенности используют три математических инструмента — вероятностно-статистический, нечеткий (основанный на теории нечетких множеств [1, разд. 11.5]), интервальный (развита статистика интервальных данных [1, гл. 12]. Авторы анализируемых методических материалов знакомы только с вероятностно-статистическим подходом [16].

Таким образом, методические материалы [13 – 18] имеют многочисленные недостатки и далеки от современного научного уровня в области статистических методов анализа данных. Это касается и других публикаций, например [20 – 22], которые также имеют различные недостатки.

Для исправления ситуации, по нашему мнению, в настоящее время необходимы прежде всего современные требования к методам обработки данных и представлению результатов статистического анализа конкретных данных. На основе

таких требований следует, в частности, разрабатывать учебники и учебные пособия. Обсудим ряд требований, о которых идет речь.

О новой парадигме прикладной статистики

Исходим из современной парадигмы прикладной статистики [23], о которой необходимо сказать несколько слов.

Статистические методы анализа данных широко применяют в различных областях науки. Обсудим смену парадигм прикладной статистики — изменения основ общепринятой модели действий в этой области математических методов исследования. Рассмотрим три реально используемых парадигмы — примитивную, устаревшую, современную.

Поясним на примере. Исходя из примитивной парадигмы, несведующие авторы применяют широко известные расчетные формулы критерия Стьюдента для проверки равенства нулю математического ожидания без какого-либо обоснования. Согласно устаревшей парадигме констатируют (без строгого обоснования), что результаты измерений имеют нормальное распределение, затем применяют критерий Стьюдента (в предположениях нормальности это обосновано). В современной парадигме используют непараметрические методы, в рассматриваемой постановке основанные на центральной предельной теореме [1].

Очевидно, обоснованность статистических выводов возрастает при переходе от примитивной парадигмы к устаревшей и далее — к современной. Констатируем, что в настоящее время в практике научной работы в различных областях используют все три парадигмы. Обсудим, как это влияет на качество результатов исследовательской деятельности.

Примитивная парадигма — это парадигма поваренной книги, т.е. следования составленным кем-то рецептам. Программные продукты часто провоцируют такие расчеты. Довольно часто итоговые выводы оказываются полезными с позиций прикладной области. Но иногда они могут быть и грубо ошибочными. Об опасности бездумного применения программных продуктов предупреждал еще проф. В. В. Налимов [24], выдающийся исследователь в области статистических методов.

Устаревшая парадигма — это парадигма середины XX в. В ней элементы выборки рассматриваются как независимые случайные величины, распределения которых входят в то или иное параметрическое семейство распределений — нормальных, логистических, экспоненциальных, Вейбулла – Гнеденко, Коши, Лапласа, гамма-распределений, бета-распределений и др. Все эти семейства входят в четырехпараметрическое семейство распределений, введенное основателем математической статистики К. Пирсоном в начале XX в. В целях упорядочения результатов измерений (наблюдений, анализов, испытаний, опытов, обследований) он принял рабочую гипотезу, что распределения реальных данных всегда совпадают с каким-то элементом его четырехпараметрического семейства. Затем началось развитие теории параметрической математической статистики, в которой задачи оценивания и проверки гипотез решались для выборок из тех или иных параметрических семейств. Был получен ряд замечательных математических моделей и результатов, например, связанных с методом максимального правдоподобия, критериями Пирсона (хи-квадрат), неравенством Рао – Крамера и др. Многомерное нормальное распределение оказалось весьма полезным для развития регрессионного и дискриминантного анализа.

Параметрической математической статистике посвящено основное содержание распространенных профильных вузовских учебников. В отличие от примитивной парадигмы, имеется строгая математическая теория, позволяющая получать расчетные алгоритмы и на их основе — полезные практические рекомендации. Есть только один недостаток — распределения реальных данных, как правило, не являются нормальными и вообще не входят в четырехпараметрическое семейство Пирсона [1]. Делают попытки проверить нормальность или, например, экспоненциальность реальных данных. Зачастую отклонить гипотезу нормальности не удается. Но это нельзя рассматривать как подтверждение нормальности распределения рассматриваемых данных, поскольку для тех же данных обычно не удается отклонить и гипотезу о том, что распределение данных соответствует другому популярному распределению. Причина очевидна — малый объем выборки. Например, для того чтобы выяснить, какому распределению соответствуют анализируемые данные — нормальному или логистическому, необходимо не менее 2500 наблюдений [1]. Реальные объемы выборок обычно значительно меньше.

Развитие теории параметрической математической статистики продолжается и в настоящее время. В частности, сравнительно недавно выяснено, что вместо оценок максимального правдоподобия целесообразно использовать одношаговые оценки, разработаны методы доверительного оценивания для параметров гамма-распределения и др. С помощью параметрической математической статистики решено много прикладных задач в конкретных областях исследования. Но в ряде случаев получены ошибочные вы-

воды, хотя доля таких случаев заметно меньше, чем при опоре на примитивную парадигму.

Современная парадигма прикладной статистики и шире — математических методов исследования [23] основана на непараметрической и нечисловой статистике. В отличие от параметрической статистики предполагается, что элементы выборки с числовыми значениями имеют произвольную непрерывную функцию распределения (во многих случаях добавляют еще условие непрерывности). Центральной областью прикладной статистики стала статистика нечисловых данных [1, главы 5, 11], позволяющая единообразно подходить к анализу статистических данных произвольной природы.

Современную парадигму математических методов исследования называем новой, хотя ее основы сформировались еще в 1980-х годах, когда во время подготовки к созданию Всесоюзной статистической ассоциации (учредительный съезд прошел в 1990 г.) понадобилось проанализировать состояние и перспективы прикладной статистики.

К настоящему времени непараметрическими методами можно решать практически тот же круг задач анализа данных, что и параметрическими. Преимущество непараметрической статистики состоит в отсутствии необходимости принимать необоснованные предположения о виде функции распределения. Недостатком является то, что реальные данные часто содержат совпадения. Если функция распределения элементов выборки непрерывна, то вероятность их совпадения равна нулю. Противоречие возникает из-за того, что свойства прагматических чисел, используемых для записи результатов измерений (наблюдений, испытаний, опытов, анализов, обследований), отличаются от свойств математических чисел, например, прагматические числа записываются с помощью конечного числа цифр, а почти все действительные числа требуют (в теории) бесконечного ряда цифр. Разработаны подходы к анализу совпадений при применении непараметрических статистик, позволяющие снять рассматриваемое противоречие [25].

Необходимо отметить, что в некоторых случаях параметрические методы позволяют обнаружить и предварительно изучить важные эффекты непараметрической статистики. Так, хорошо известно, что распределения реальных данных, как правило, не являются нормальными. Однако математический аппарат в случае нормальности зачастую является более простым. Согласно устаревшей парадигме в математической статистике широко используют многомерные нормальные распределения. Именно для таких распределений найдены явные формулы для различных характеристик в многомерном статистическом анализе, прежде всего в регрессионных постановках. Это связано с тем, что глубоко развита теория квадратичных форм в евклидовом пространстве (квадратичные формы стоят в степени экспоненты, описывающей плотность многомерного нормального распределения). Используя развитый математический аппарат, основанный на многомерной нормальности, удается, например, разработать и изучить методы оценивания размерности вероятностно-статистической модели [1, гл. 9] в целях переноса полученных результатов на непараметрические постановки.

В настоящее время теоретические исследования по прикладной статистике проводят в основном в соответствии с современной парадигмой. Так, статистике нечисловых данных посвящено 63 % работ по прикладной статистике, опубликованных в разделе «Математические методы исследования» журнала «Заводская лаборатория. Диагностика материалов» в 2006 – 2015 гг. Однако значительная доля прикладных работ осуществляется в традициях устаревшей или даже примитивной парадигмы. Такие работы нецелесообразно огульно отрицать. Они могут приносить пользу в конкретных областях. Однако бесспорно, что переход на современную парадигму прикладной статистики повысит научный уровень исследований, а также позволит получить важные результаты в конкретных областях. Приходится констатировать, что исследователи, связанные с анализом данных, недостаточно знакомы с непараметрической и нечисловой статистикой. Необходимо шире распространять информацию о современной парадигме прикладной статистики.

Опора на подходы и результаты непараметрической и нечисловой статистики — одно из основных требований к статистическим методам анализа данных.

Вероятностно-статистические модели данных — основа методов прикладной статистики

При описании и обсуждении процедур анализа статистических данных обычно сосредотачивают внимание на расчетных формулах. Причина очевидна — не зная формул, нельзя провести расчеты. Однако начинать надо с вероятностно-статистических моделей порождения изучаемых данных.

Например, в прикладной статистике наиболее распространенная модель выборки — это конечная последовательность независимых одинаково распределенных случайных величин [1], моделирующих результаты измерений (наблюдений, испытаний, опытов, анализов, обследований). Если общая функция распределения этих

случайных величин является произвольной, то необходимо обратиться к методам непараметрической статистики. Для реальных данных совпадения результатов встречаются достаточно часто. Следовательно, в таких случаях наблюдаются отклонения от непараметрической модели. Как отмечено выше, модель анализа совпадений при расчете непараметрических ранговых статистик предложена в работе [25]. Статистика интервальных данных была создана для обработки округленных данных и данных с совпадениями [1, гл. 12].

Отметим устойчивость предрассудков. Например, до сих пор пропагандируется использование метода максимального правдоподобия, хотя свойства одношаговых оценок не менее хороши. Однако во многих случаях система уравнений максимального правдоподобия не имеет явного решения и соответствующие оценки рекомендуется находить итерационными методами, сходимость которых не изучают, хотя есть примеры, в которых отсутствие сходимости продемонстрировано. Между тем одношаговые оценки вычисляют по конечным формулам, без всяких итераций [1].

Особенно заметна любовь теоретиков в области прикладной статистики к многомерным нормальным распределениям. Именно для таких распределений найдены явные формулы для различных характеристик в многомерном статистическом анализе, прежде всего в регрессионном. По нашей экспертной оценке, причина в том, что удается использовать хорошо развитую в линейной алгебре теорию квадратичных форм.

Известно, что распределения почти всех реальных данных ненормальны. Это утверждение хорошо обосновано экспериментально путем анализа результатов измерений [1]. Теоретические аргументы в пользу нормального распределения также не выдерживают критики. Например, говорят, что зависимость значения случайной величины от многих факторов влечет нормальность. Иногда добавляют, что факторы являются независимыми и сравнимыми по величине. Однако нормальность распределения можно ожидать лишь в случае аддитивной модели, когда факторы складываются (в силу Центральной предельной теоремы). Если же случайная величина формируется путем перемножения (мультипликативная модель), то ее распределение является (в асимптотике) логарифмически нормальным. Если справедлива модель «самого слабого» звена (или «самого сильного»), т.е. случайная величина равна крайнему члену вариационного ряда значений факторов (соответственно минимуму или максимуму), то имеем в пределе распределение Вейбулла – Гнеденко.

Модель на основе семейства нормальных распределений или распределений из иного параметрического семейства можно сравнить с моделью поиска под фонарем потерянных в темных кустах ключей. Очевидно, под фонарем искать легче. Можно продемонстрировать активность, но надеяться на благоприятный исход поисков нельзя.

Из проведенного анализа следует необходимость использования непараметрических моделей распределений результатов измерений. Отметим, что интервалы их возможных значений, как правило, ограничены, т.е. распределения являются финитными. Следовательно, все моменты рассматриваемых случайных величин существуют и их выборочные аналоги могут использоваться в вычислениях.

Сформулируем вытекающее из сказанного требование к статистическим методам обработки данных: если по каким-либо причинам применяется параметрическое семейство распределений, его использование должно быть тщательно обосновано путем проверки гипотезы согласия как с рассматриваемым семейством, так и с альтернативными семействами.

Роль вероятностно-статистических моделей в многомерном статистическом анализе

Начнем с регрессионного анализа. Используют четыре основных класса регрессионных моделей.

К первому типу отнесем распространенные модели метода наименьших квадратов с детерминированной независимой переменной и параметрической зависимостью (линейной, квадратической и т.п.). Распределение отклонений произвольно (т.е. модель является непараметрической), для получения предельных распределений оценок параметров и регрессионной зависимости предполагаем выполнение условий центральной предельной теоремы.

Второй тип моделей основан на выборке случайных векторов. Зависимость является параметрической, распределение двумерного вектора — произвольным. Об оценке дисперсии независимой переменной можно говорить только в модели на основе выборки случайных векторов, равно как и о коэффициенте детерминации как критерии качества модели.

Третий тип моделей регрессионного анализа, основанный на выборке случайных векторов, — непараметрическая регрессия, в которой как зависимость, так и отклонения от нее являются непараметрическими. Зависимость (как условное среднее) оценивается с помощью непараметрических оценок плотности. Промежуточный вари-

ант — модель, в которой тренд линеен, а периодическая и случайная составляющие и отклонения от них являются непараметрическими.

В моделях четвертого типа малые погрешности имеются в значениях как зависимой, так и независимой переменных. В прошлом этот раздел прикладной статистики назывался конфлюэнтным анализом, сейчас он входит в статистику интервальных данных [1, гл. 12].

К регрессионному анализу примыкают задачи сглаживания временных рядов и статистики случайных процессов, в которых отклонения от функции времени зависимы (в отличие от регрессионного анализа, в котором такие отклонения — независимые случайные величины).

Анализ многообразия моделей регрессионного анализа приводит к выводу, что не существует единой «стандартной модели». Другими словами, при решении задачи восстановления зависимости необходимо начинать с выбора и обоснования той или иной вероятностно-статистической модели.

Теория измерений как основа построения вероятностно-статистических моделей

Необходимо исходить из теории измерений [1, гл. 5], согласно которой первый шаг при анализе данных — выявление шкал, в которых они измерены. Применяемые статистические методы должны соответствовать шкалам, в которых измерены данные.

Так, известно, что для данных, измеренных в порядковой шкале, в качестве средних величин можно использовать только члены вариационного ряда, прежде всего медиану при нечетном объеме данных, а при четном — левую медиану или правую медиану. Применение среднего арифметического или среднего геометрического недопустимо. Поскольку ранги или баллы, как правило, измерены в порядковой шкале, складывать их нельзя. В частности, нельзя оценивать успеваемость учащихся по среднему баллу экзаменационных оценок.

Статистические выводы должны быть инвариантны относительно допустимых преобразований шкал измерения данных. Значит, для каждой шкалы можно выяснить, какими алгоритмами анализа данных из рассматриваемого семейства можно пользоваться в этой шкале. Выше описаны выводы относительно семейства средних по Коши. Обратная задача — для определенного алгоритма анализа данных выяснить, в какой шкале можно им пользоваться. Коэффициент линейной парной корреляции Пирсона соответствует шкале интервалов, а непараметрические ранговые коэффициенты корреляции Спирмена и Кендалла позволяют изучать взаимосвязи порядковых переменных.

С позиций теории измерений обсудим довольно широко известный метод анализа иерархий. Исходные данные — результаты парных сравнений, они измерены в порядковых шкалах. А результаты расчетов методом анализа иерархий выражены в шкале интервалов. С точки зрения теории измерений такое недопустимо. Следовательно, методом анализа иерархий пользоваться не следует. Рекомендуем применять адекватные методы анализа экспертных оценок, в частности, методы средних арифметических рангов, медиан рангов, согласования кластеризованных ранжировок [26].

Теория классификации, обучающие выборки и нейросети

Вполне естественно распространить разрабатываемые требования на смежную (близкородственную) область — нейросетевую обработку данных. Учитывая значительное взаимопроникновение вероятностно-статистических и нейросетевых методов, это представляется весьма полезно.

В целях реализации этой идеи рассмотрим конкретную область прикладной статистики — теорию классификации. В ней выделяют три области — построение классификаций, изучение классификаций, применение классификаций [1, гл. 9]. Если изучение классификаций — это однозначно часть статистики нечисловых данных, то две другие области имеют в литературе самые разные названия.

Построение классификаций — это то же самое, что кластер-анализ (кластерный анализ), распознавание образов без учителя, типология, таксономия, группировка, классификация без учителя, дихотомия…

Применение классификаций — это то же самое, что дискриминация (дискриминантный анализ), диагностика, распознавание образов с учителем, автоматическая классификация с учителем, статистическая классификация...

«Учитель», о котором здесь идет речь, — это использование обучающих выборок. В этом случае классы заданы обучающими выборками, и на их основе формируется правило принятия решений о том, к какому классу отнести вновь поступающий объект. Когда говорят об алгоритмах без учителя, речь идет о построении классификации на основе анализа данных единой обучающей выборки.

В настоящее время популярный термин — нейросети. Речь идет о математических моделях (и разработанных на их основе программной или аппаратной реализациях), построенных по ана-

логии с сетями нервных клеток живого организма. Эти модели возникли в середине XX в. при изучении процессов, протекающих в мозге, и при попытке смоделировать эти процессы (на уровне знаний того времени).

Если вникнуть в суть нейросетевых методов, то становится очевидным, что эти модели предназначены прежде всего для решения задач классификации на основе анализа обучающих выборок. При этом нейросетевые алгоритмы, как правило, не являются оптимальными. Например, доказано, что для отнесения вновь поступающего объекта в один из двух классов, заданных обучающими выборками, (асимптотически) оптимальным является решающее правило, основанное на непараметрических оценках плотностей распределений вероятностей, соответствующих классам [1]. Нейросетевые методы не могут дать лучшего результата, чем это решающее правило. Однако частое упоминание нейросетей в современной литературе приводит к забвению оптимальных методов и алгоритмов, что, естественно, снижает эффективность технологических решений искусственного интеллекта.

Констатируем, что нейросети, методы распознавания образов и, например, генетические алгоритмы — это другие названия ряда давно разрабатываемых разделов прикладной статистики (статистических методов анализа данных) [1]. Новая терминология вынесена на передний план внимания научной общественности по вненаучным причинам.

Выводы

Как следует из сказанного выше, необходима разработка системы требований к статистическим моделям и методам при их создании, применении и преподавании, в том числе при их описании в публикациях.

Прежде всего, должна быть представлена и обоснована вероятностно-статистическая модель порождения данных. Иерархия понятия «модель» и потенциальные источники ошибок при построении вероятностно-статистической модели реальных данных исследованы в [27].

Приведем ряд требований к статистическим методам, проанализированных выше.

Поскольку практически все распределения реальных данных ненормальны, предпочтения следует отдавать непараметрическим постановкам. Возможность применения параметрических семейств распределений должна быть тщательно обоснована.

В соответствии с теорией проверки статистических гипотез должны быть указаны не только нулевая гипотеза, но и альтернативная, только тогда можно обсуждать мощность критерия.

Необходимо изучение устойчивости выводов, получаемых на основе модели, относительно допустимых изменений исходных данных и предпосылок модели [1, разд. 4.7]. В частности, статистические выводы должны быть инвариантны относительно допустимых преобразований шкал.

Проблемам разработки системы требований к статистическим моделям и методам будет посвящен ряд дальнейших публикаций.

ЛИТЕРАТУРА

1. **Орлов А. И.** Прикладной статистический анализ. — М.: Ай Пи Ар Медиа, 2022. — 812 с. DOI: 10.23682/117038
2. **Orlov A. I.** Basic requirements for statistical methods of data analysis / Polythematic Online Scientific Journal of Kuban State Agrarian University. 2022. N 181. P. 316 – 343. EDN OKGBOS. DOI: 10.21515/1990-4665-181-026
3. **Бурбаки Н.** Очерки по истории математики. — М.: Изд-во иностранной литературы, 1963. — 292 с.
4. **Колмогоров А. Н.** Основные понятия теории вероятностей. — Изд. 2-е. — М.: Наука, 1974. — 120 с.
5. **Крамер Г.** Математические методы статистики. — Изд. 2-е. — М.: Мир, 1975. — 648 с.
6. **Большев Л. Н., Смирнов Н. В.** Таблицы математической статистики. — Изд. 3-е. — М.: Наука. 1983. — 416 с.
7. Вероятность и математическая статистика: Энциклопедия / Гл. ред. акад. Ю. В. Прохоров. — М.: Большая Рос. Энцикл., 1999. — 910 с.
8. **Боровков А. А.** Математическая статистика. — Изд. 5-е, стер. — Санкт-Петербург: Лань, 2021. — 704 с.
9. **Боровков А. А.** Математическая статистика. Дополнительные главы. — М.: Наука, 1984. — 144 с.
10. **Чибисов Д. М., Пагурова В. И.** Задачи по математической статистике. — М.: Изд-во Московского университета, 1990. — 171 с.
11. **Смирнов Н. В., Дунин-Барковский И. В.** Курс теории вероятностей и математической статистики для технических приложений. — М.: Наука, 1969. — 512 с.
12. **Ивченко Г. И., Медведев Ю. И.** Математическая статистика: Учебник. — М.: Книжный дом «ЛИБРОКОМ», 2014. — 352 с.
13. **Волковец А. И., Гуринович А. Б.** Теория вероятностей и математическая статистика. Конспект лекций. — Минск: БГУИР, 2003. — 84 с.
14. **Губарь Л. Н., Ермоленко А. В.** Теория вероятностей и математическая статистика: учебное пособие. — Сыктывкар: Изд-во СГУ имени Питирима Сорокина, 2015. — 120 с.
15. **Кибзун А. И., Горяинова Е. Р., Наумов А. В., Сиротин А. Н.** Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами. — М.: Физматлит, 2002. — 224 с.
16. **Симонова И. Э., Симонов А. Б., Сагателова Л. С.** Теория вероятностей и математическая статистика. — Волгоград: Волгоградский государственный технический университет, 2022. — 96 с.
17. **Трофимова Е. А., Кисляк Н. В., Гилёв Д. В.** Теория вероятностей и математическая статистика: учеб. пособие. — Екатеринбург: Изд-во Урал. ун-та, 2018. — 160 с.
18. **Чебоксаров А. Б., Иванова И. Б.** Теория вероятностей и математическая статистика: учебное пособие. — Пятигорск: ООО «Рекламно-информационное агентство на КМВ», 2020. — 80 с.
19. **Налимов В. В.** Применение математической статистики при анализе вещества. — М.: Физматлит, 1960. — 430 с.
20. **Гмурман В. Е.** Теория вероятностей и математическая статистика: учебник для вузов. Изд. 12-е. — М.: ЮРАЙТ, 2021. — 479 с.
21. **Кремер Н. Ш.** Теория вероятностей и математическая статистика: учебник. — М.: Юнити, 2012. — 551 с.

22. **Кобзарь А. И.** Прикладная математическая статистика: для инженеров и научных работников. — М.: Физматлит, 2006. — 816 с.
23. **Орлов А. И.** Новая парадигма прикладной статистики / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 1. С. 87 – 93.
24. **Налимов В. В.** Теория эксперимента. — М.: Наука, 1971. — 208 с.
25. **Орлов А. И.** Модель анализа совпадений при расчете непараметрических ранговых статистик / Заводская лаборатория. Диагностика материалов. 2017. Т. 83. № 11. С. 66 – 72. DOI: 10.26896/1028-6861-2017-83-11-66-72
26. **Орлов А. И.** Искусственный интеллект: экспертные оценки. — М.: Ай Пи Ар Медиа, 2022. — 436 с. DOI: 10.23682/117030
27. **Савельев О. Ю.** Модель: иерархия понятия и потенциальный источник ошибок / Инновации в менеджменте. 2021. № 28. С. 54 – 58.

REFERENCES

1. **Orlov A. I.** Applied Statistical Analysis. — Moscow: Ai Pi Ar Media, 2022. — 812 p. [in Russian]. DOI: 10.23682/117038
2. **Orlov A. I.** Basic requirements for statistical methods of data analysis / Polythematic Online Scientific Journal of Kuban State Agrarian University. 2022. N 181. P. 316 – 343. EDN OKGBOS. DOI: 10.21515/1990-4665-181-026
3. **Burbaki N.** Essays on the history of mathematics. — Moscow: Izd. inostrannoi literatury, 1963. — 292 p. [in Russian].
4. **Kolmogorov A. N.** Basic concepts of probability theory. 2nd ed. — Moscow: Nauka, 1974. — 120 p. [in Russian].
5. **Kramer G.** Mathematical methods of statistics. 2nd ed. — Moscow: Mir, 1975. — 648 p. [Russian translation].
6. **Bol'shev L. N., Smirnov N. V.** Tables of mathematical statistics. 3rd ed. — Moscow: Nauka. 1983. — 416 p. [in Russian].
7. **Prokhorov Yu. V., Ed.** Probability and Mathematical Statistics: Encyclopedia. — Moscow: Bol'shaya Ros. Éntsiklopediya, 1999. — 910 p. [in Russian].
8. **Borovkov A. A.** Mathematical statistics. 5th ed. — St. Petersburg: Lan', 2021. — 704 p. [in Russian].
9. **Borovkov A. A.** Mathematical statistics. Additional chapters. — Moscow: Nauka, 1984. — 144 p. [in Russian].
10. **Chibisov D. M., Pagurova V. I.** Problems in mathematical statistics. — Moscow: Izd. MGU, 1990. — 171 p. [in Russian].
11. **Smirnov N. V., Dunin-Barkovskii I. V.** Course of Probability Theory and Mathematical Statistics for Technical Applications. — Moscow: Nauka, 1969. — 512 p. [in Russian].
12. **Ivchenko G. I., Medvedev Yu. I.** Mathematical Statistics: Textbook. — Moscow: Knizhnyi dom "Librokom", 2014. — 352 p. [in Russian].
13. **Volkovets A. I., Gurinovich A. B.** Theory of Probability and Mathematical Statistics. Lecture notes. — Minsk.: BGUIR, 2003. — 84 p. [in Russian].
14. **Gubar' L. N., Ermolenko A. V.** Theory of Probability and Mathematical Statistics: Tutorial. — Syktyvkar: Izd. SGU imeni Pitirima Sorokina, 2015. — 120 p. [in Russian].
15. **Kibzun A. I., Goryainova E. R., Naumov A. V., Sirotin A. N.** Theory of Probability and Mathematical Statistics. Basic course with examples and tasks. — Moscow: Fizmatlit, 2002. — 224 p. [in Russian].
16. **Simonova I. E., Simonov A. B., Sagatelova L. S.** Theory of Probability and Mathematical Statistics. — Volgograd: VGTU, 2022. — 96 p. [in Russian].
17. **Trofimova E. A., Kislyak N. V., Gilyov D. V.** Theory of Probability and Mathematical Statistics: Tutorial. — Yekaterinburg: Izd. UrFU, 2018. — 160 p. [in Russian].
18. **Cheboksarov A. B., Ivanova I. B.** Theory of Probability and Mathematical Statistics: Tutorial. — Pyatigorsk: OOO "Reklamno-informatsionnoe agentstvo na KMV", 2020. — 80 p. [in Russian].
19. **Nalimov V. V.** Application of mathematical statistics in the analysis of substances. — Moscow: Fizmatlit, 1960. — 430 p. [in Russian].
20. **Gmurman V. E.** Probability theory and mathematical statistics: a textbook for universities. 12th ed. — Moscow: Yurait, 2021. — 479 p. [in Russian].
21. **Kremer N. Sh.** Probability Theory and Mathematical Statistics: textbook. — Moscow: Yuniti, 2012. — 551 p. [in Russian].
22. **Kobzar' A. I.** Applied Mathematical Statistics: for engineers and scientists. — Moscow: Fizmatlit, 2006. — 816 p. [in Russian].
23. **Orlov A. I.** The New Paradigm of Applied Statistics / Industr. Lab. Mater. Diagn. 2012. Vol. 78. N 1. P. 87 – 93 [in Russian].
24. **Nalimov V. V.** Theory of experiment. — Moscow: Nauka, 1971. — 208 p. [in Russian].
25. **Orlov A. I.** The model of coincidence analysis in the calculation of nonparametric rank statistics / Industr. Lab. Mater. Diagn. 2017. Vol. 83. N 11. P. 66 – 72 [in Russian]. DOI: 10.26896/1028-6861-2017-83-11-66-72
26. **Orlov A. I.** Artificial intelligence: expert estimations. — Moscow: Ai Pi Ar Media, 2022. — 436 p. [in Russian]. DOI: 10.23682/117030
27. **Savel'ev O. Yu.** Model: Concept Hierarchy and Potential Error Source / Innov. Menedzh. 2021. N 28. P. 54 – 58 [in Russian].