

Математические методы исследования

Mathematical methods of investigation

DOI: <https://doi.org/10.26896/1028-6861-2020-86-11-67-78>

ОСНОВНЫЕ ТРЕБОВАНИЯ К МАТЕМАТИЧЕСКИМ МЕТОДАМ КЛАССИФИКАЦИИ

© 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 21 января 2020 г. Поступила после доработки 21 января 2020 г.
Принята к публикации 10 февраля 2020 г.*

Упорядочивание методов классификации должно повысить их роль в решении прикладных задач, в частности, при диагностике материалов. Для этого прежде всего следует выработать требования, которым должны удовлетворять методы классификации. Первоначальная формулировка таких требований — основное содержание настоящей работы. Математические методы классификации рассматриваются как часть методов прикладной статистики. Обсуждаются естественные требования к рассматриваемым методам анализа данных и представлению результатов расчетов, вытекающие из накопленных отечественной вероятностно-статистической научной школой достижений и идей. Приведены конкретные рекомендации по ряду вопросов, а также критика отдельных ошибок. В частности, методы анализа данных должны быть инвариантны относительно допустимых преобразований шкал, в которых измерены данные, т.е. методы должны быть адекватны в смысле теории измерений. Основой конкретного статистического метода анализа данных всегда является та или иная вероятностная модель. Она должна быть явно описана, ее предпосылки обоснованы — либо из теоретических соображений, либо экспериментально. Методы обработки данных, предназначенные для использования в реальных задачах, должны быть исследованы на устойчивость относительно допустимых отклонений исходных данных и предпосылок модели. Необходимо знать точность решений, получаемых с помощью используемого метода. При публикации результатов статистического анализа реальных данных следует указывать их точность (доверительные интервалы). В качестве оценки прогностической силы алгоритма классификации вместо доли правильных прогнозов рекомендуется использовать прогностическую силу. Математические методы исследования делятся на «разведочный анализ» и «доказательную статистику». Специфические требования к методам обработки данных возникают в связи с их «стыковкой» при последовательном выполнении. Обсуждаются границы применимости вероятностно-статистических методов. Рассмотрены также конкретные постановки задач классификации и типовые ошибки при применении различных методов их решения.

Ключевые слова: прикладная статистика; анализ данных; методы классификации; диагностика; теория измерений; устойчивость; прогностическая сила.

BASIC REQUIREMENTS FOR MATHEMATICAL METHODS OF CLASSIFICATION

© Alexander I. Orlov

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

Received January 21, 2020. Revised January 21, 2020. Accepted February 10, 2020.

Ordering of classification methods should increase their role in solving applied problems, in particular, in diagnostics of materials. Development of the requirements to be met by classification methods is the first concern. The initial formulation of the requirements is the main content of this work. Mathematical methods of classification are considered as a part of the applied statistics methods. The natural requirements to the considered methods of data analysis and presentation of the calculation results arising from the achievements and ideas accumulated by the national probabilistic and statistical scientific school are discussed. Specific recommendations are given on a number of issues, as well as criticism of individual

errors. In particular, methods of data analysis must be invariant with respect to the permissible transformations of the scales in which the measured data are gained, i.e., the methods should be adequate in the terms of measurement theory. The basis of any statistical method of data analysis is always a particular probabilistic model which must be clearly described and the premises must be justified either from theoretical considerations, or experimentally. Data processing methods intended for use in real-world problems should be tested for robustness with respect to the acceptable deviations of the initial data and model assumptions. The accuracy of the solutions provided by the method should be determined. When publishing the results of statistical analysis of real data, it is necessary to indicate their accuracy (confidence intervals). As an estimate of the predictive power of the classification algorithm, it is recommended to use predictive power instead of the proportion of correct forecasts. Methods of mathematical research are divided into “exploratory analysis” and “evidence-based statistics.” Specific requirements for data processing methods arise in connection with their “docking” during their sequential execution. The limits of applicability of probabilistic-statistical methods are discussed. Specific statements of the classification problems and typical errors in application of different methods of their solution are also considered.

Keywords: applied statistics; data analysis; classification methods; diagnostics; measurement theory; stability; predictive power.

Введение

Во всех отраслях промышленности, в медицине, социально-экономических исследованиях и других областях деятельности постоянно решаются разнообразные задачи классификации. Разработано много различных математических методов классификации. Строго говоря, их не меньше, чем точек на отрезке. Действительно, ряд методов использует только расстояния между классифицируемыми объектами. Однако, если d — расстояние (метрика), то d^α — также метрика при любом α таком, что $0 < \alpha < 1$.

Несмотря на многообразие постановок задач, моделей и методов классификации, алгоритмов расчетов, положение дел в этой области анализа данных далеко от удовлетворительного. Задачи классификации зачастую решаются не наилучшим образом (более того, зачастую не ясно, как сравнивать методы решения). Области применимости различных методов классификации не установлены, свойства методов недостаточно изучены. Отдельные группы специалистов (кланы) разрабатывают собственные подходы, не слишком интересуясь результатами других. Популярность тех или иных методов зачастую определяется субъективными причинами. Распространен ряд сомнительных концепций и попросту заблуждений. Во многом трудности определяются тем, что накоплено столько теоретических и практических разработок, что отдельный специалист или небольшая группа не в состоянии их осмыслить.

Назрела необходимость навести порядок в методах классификации. Это повысит их роль в решении прикладных задач, в частности, при диагностике материалов. Решить поставленную задачу можно только с помощью добровольной стандартизации. Необходимо проанализировать накопленное и разработать стандарты (предприятий и организаций) по применению признанных наилучшими методов классификации. Для

этого прежде всего следует выработать требования, которым должны удовлетворять методы классификации. Первоначальная формулировка таких требований — основное содержание настоящей работы.

Методы классификации рассматриваем как часть прикладной статистики. Ниже приведем ряд примеров нарушения обсуждаемых требований к методам анализа данных, при этом критика конкретной публикации не означает, что в ней нет ничего ценного.

Требования к методам анализа данных и представлению результатов расчетов

1. Методы должны быть объективными, результат их применения — определяться исходными данными, но не субъективными мнениями и решениями исследователя. В частности, *методы анализа данных должны быть инвариантны относительно допустимых преобразований шкал, в которых измерены данные, т.е. методы должны быть адекватны в смысле теории измерений* [1]. Это требование иногда бывает довольно жестким. Так, в качестве средних величин для данных, измеренных в порядковой шкале, можно использовать только члены вариационного ряда, в частности, медиану, но не среднее арифметическое, среднее гармоническое и т.д. Из всех средних по Колмогорову условие адекватности выделяет для данных, измеренных в интервальной шкале, только среднее арифметическое, а для шкалы отношений — только степенные средние [2].

Иногда грациям порядковых данных пытаются приписать числа, чтобы потом применять методы, разработанные для количественных шкал. Это — так называемая «оцифровка» [3, 4]. Она частично оправдана лишь в том случае, когда есть уверенность, что наблюдаемые данные получены в результате группировки количественных переменных. Пропаганда методов

«оцифровки» вне указанных пределов может привести к неадекватным рекомендациям и повлечь те или иные потери. Примером неадекватной оцифровки является метод анализа иерархий [5], в котором от порядковых переменных осуществляется переход к измерениям в шкале интервалов.

2. Основой конкретного статистического метода анализа данных всегда является та или иная вероятностная модель. Именно на основе модели осуществляется перенос выводов с выборочной совокупности на более широкую (генеральную) совокупность. Модель должна быть явно описана, ее предпосылки обоснованы — либо из теоретических соображений, либо экспериментально. Так, в математической статистике часто предполагается, что данные представляют собой выборку, т.е. моделируются как реализации набора независимых одинаково распределенных случайных величин. В обосновании нуждаются, в частности, независимость, одинаковая распределенность. Обоснование используемой модели может быть дано либо из содержательных соображений (например, на основе анализа условий наблюдений), либо же путем статистической проверки. Так, критерии независимости результатов наблюдений приведены в [6, 7]. Иногда высказываемое мнение [8], что положениям математической статистики не угрожает опытная проверка, не соответствует действительности. Построением вероятностно-статистических моделей в связи с задачами классификации занимался Л. Г. Малиновский [9].

Модель и метод (алгоритм) — две самостоятельные составляющие процедуры анализа данных. Для одной и той же модели могут быть предложены различные алгоритмы. Например, параметры функции распределения можно оценивать методом моментов, методом максимального правдоподобия и др. Отметим здесь, что итеративные процедуры нахождения оценок максимального правдоподобия применять нецелесообразно: если эти оценки нельзя найти явно, то следует вычислять одношаговые оценки [1].

Более важно, что один и тот же алгоритм в одной модели может быть наилучшим из возможных, а в другой — очень плохим. Так, для проверки однородности двух выборок в классической модели, в которой элементы выборки имеют нормальные распределения, критерий Стьюдента является наилучшим (при условии равенства дисперсий). Если же распределения, из которых взяты выборки, могут быть произвольными, то этот критерий несостоятелен. К сожалению, неправильное понимание критерия Стьюдента укоренилось, например, в медицинской науке. Следует, конечно, переучивать прикладников на непараметрические критерии.

Полезно сказать несколько слов в защиту критерия Стьюдента. Во-первых, распределение статистики Стьюдента устойчиво к малым отклонениям от нормальности. Во-вторых, из Центральной предельной теоремы следует, что статистика Стьюдента распределена асимптотически нормально, если объемы обеих выборок стремятся к бесконечности, а распределения, из которых они взяты, имеют дисперсии. Отсюда следует, что критерий Стьюдента является состоятельным для проверки гипотезы о равенстве математических ожиданий двух распределений. Если последняя гипотеза отвергнута, то однородности нет (подробности см. в [10]).

Аналогичное замечание можно сделать по поводу распространенного неправильного мнения о том, что проверять равенство нулю линейного парного коэффициента корреляции Пирсона можно только в случае, когда результаты наблюдений имеют двумерное нормальное распределение. На самом же деле выборочный коэффициент корреляции асимптотически нормален, а потому при большом объеме выборки можно пользоваться теми же процедурами, что и в предположении нормальности [1].

Проверка однородности — одна из процедур классификации. А именно проверяется, представляют ли выборки два класса или же их можно объединить в один. Каким же непараметрическим критерием пользоваться? В литературе имеется много предложений, например, в [7] предлагается применять критерий Вилкоксона. Эта рекомендация не соответствует традициям отечественной вероятностно-статистической научной школы [11], рекомендующей критерии, основанные на эмпирических функциях распределения. Обсудим обоснованность рекомендации по применению критерия Вилкоксона.

В [7] критерий Вилкоксона опирается на модель, в которой одна из функций распределения произвольна, а вторая отличается от нее только сдвигом. Редко можно указать ситуацию, в которой подобная модель обоснована. Разве что при анализе результатов многократных измерений значений физической величины для двух образцов с помощью одного и того же средства измерения, для которого характеристики погрешностей стабильны в рассматриваемом диапазоне.

Если реальная ситуация достаточно изучена, то функции распределения в основном известны. Под таким заявлением обычно понимают то, что они известны с точностью до параметров, а тогда проверка гипотезы однородности проводится с помощью параметрических критериев, в частности, при нормальных распределениях с одинаковыми дисперсиями — с помощью критерия Стьюдента.

Когда реальная ситуация изучена мало, функции распределения естественно считать произвольными и не связанными друг с другом. Затруднительно представить себе ситуацию, в которой связь между функциями распределения известна почти полностью (с точностью до параметра сдвига), в то время как о самих функциях распределения ничего не известно. Авторы [7] не рассматривают такие ситуации, в соответствующем примере [7, с. 87 – 88] они попросту не обосновывают модель. Таким образом, не сведущий в прикладной статистике исследователь, пользуясь [7], может взять произвольную модель, обработать данные в соответствии с ней, результат расчетов выдать как научно обоснованный.

(Отметим, что название [7] не соответствует содержанию: эту монографию следовало бы назвать «Избранные ранговые статистические методы». В [7] несколько искажена история непараметрической статистики, полностью игнорируются такие ее современные разделы, как непараметрические оценки плотности и регрессии. Современные взгляды на непараметрическую статистику обсуждаются в статье [12].)

Итак, при проверке однородности в непараметрическом случае необходимо принять, что функции распределения выборок произвольны. В такой постановке критерий Вилкоксона не является состоятельным. Значит, его применять нельзя. Чем же пользоваться? Очевидно, состоятельными критериями — Смирнова, типа омега-квадрат (Лемана – Розенблатта) [11] и др. А каким именно — нерешенная проблема, подходов к которой не видно (она стоит первой в «цахкадзорской тетради» [13]). Если известна альтернатива, то можно подобрать наиболее мощный критерий. Но откуда взять альтернативу?

Ясно, что нельзя ждать, пока наука дозреет до решения этой проблемы. В настоящее время мы считаем целесообразным рекомендовать два критерия — двухсторонний критерий Смирнова и типа омега-квадрат (Лемана – Розенблатта). В пользу первого из них говорит то, что разработан быстрый алгоритм вычисления распределения критерия Смирнова при конечных объемах выборок, на основе которого рассчитаны таблицы критических значений, исчерпывающим образом дополняющие таблицы для предельного распределения [14]. (Отметим, что называть этот критерий «критерием Колмогорова – Смирнова», как это сделано в [7], неправильно, поскольку у Колмогорова и Смирнова не было ни одной совместной работы. Рассматриваемый критерий был предложен Н. В. Смирновым в 1939 г., причем, вопреки [7], метод нахождения предельного распределения статистики Смирнова никак не связан с методом известной работы А. Н. Колмогорова 1933 г., в которой введен «критерий Кол-

могорова».) Однако у критерия Смирнова имеется заметный недостаток — его функция распределения растет большими скачками, а потому реальный уровень значимости может сильно отличаться от номинального [15]. Поэтому в настоящее время [16] мы склоняемся к использованию критерия типа омега-квадрат (Лемана – Розенблатта).

Приведенная выше критика критерия Вилкоксона относится также и к его обобщениям, применяемым в так называемом «непараметрическом дисперсионном анализе» [7] (кстати, название это неточно, поскольку никаких «дисперсий» в данных непараметрических методах нет). В рассматриваемых постановках также необходимо перейти на состоятельные критерии.

Таким образом, на примере проверки гипотезы однородности показаны необходимость обоснования вероятностной модели реального явления и ее взаимосвязь с алгоритмом расчетов, а также продемонстрирован ряд типичных ошибок.

3. Методы обработки данных, предназначенные для использования в реальных задачах, должны быть исследованы на устойчивость относительно допустимых отклонений исходных данных и предпосылок модели. В частности, следует указывать точность решений, определяемую по точности исходных данных. При этом каждый отдельный элемент исходных данных (например, элемент выборки) надо рассматривать как представитель кластера, сгустка с размытыми границами, определяемыми погрешностями исходных данных. Решения, даваемые моделью, описываются, естественно, как элементы кластера — образа кластера данных. Этот подход подробно рассмотрен в монографиях [17, 18], а применительно к теории классификации — в [19] и других статьях. Здесь отметим только два применения развитой нами общей теории устойчивости.

Анализ погрешностей социологических данных позволяет заключить, что в социологических (и маркетинговых) анкетах не имеет смысла использовать более трех – шести градаций [17, п. 2.6]. Различие значений параметров моделей управления запасами, определяемых по методикам тех или иных организаций, приводило отдельных экономистов к выводу о невозможности использования оптимизационных моделей. Анализ с позиций теории устойчивости показал, что все рассматриваемые значения лежат в одном и том же кластере, определяемом погрешностями, а анализ кластера решений привел к выводу, что оптимизационная модель позволяет снизить издержки не менее чем в два раза [17, п. 5.1].

Заслуживает дальнейшего развития связь разработанной нами теории устойчивости с теорией решения некорректных задач [20] и с теори-

ей нечеткости. Отметим, что в [17, 18] указан способ сведения теории нечеткости к теории случайных множеств, что позволяет рассматривать теорию нечеткости как своеобразный частный вероятностно-статистический метод. Ясно также, что нечеткость границ реально существующих кластеров должна учитываться в алгоритмах кластер-анализа, т.е. во многих реальных задачах адекватной является лишь нечеткая классификация.

4. Необходимо указывать точность решений, получаемых с помощью используемого метода. Понятие «точность» конкретизируется для отдельных классов методов. Так, погрешности решения могут быть связаны с погрешностями исходных данных, с погрешностями округления при компьютерных вычислениях, с погрешностями выбранного численного метода решения строго поставленной математической задачи, с тем, что математическая модель лишь грубо отражает действительность, и т.д. Особенно важно уметь численно оценивать погрешности при использовании так называемых «эвристических» алгоритмов, таких как алгоритм [21], о котором авторы честно пишут, что не знают, дает ли он решение поставленной оптимизационной задачи.

Надо констатировать, что каждый метод обработки данных — это косвенное измерение [1, 17, 18]. Перед массовым использованием, как и всякий метод измерения, он должен быть обоснован с позиций метрологии (науки об измерениях). Поскольку аналитические методы при конечных объемах выборок зачастую не разработаны, то напрашивается изучение точности решений с помощью метода Монте-Карло. Однако следует знать, что многие используемые ныне датчики псевдослучайных чисел дают последовательности, свойства которых явно отличаются от номинальных при числе испытаний, скажем, более 2000, как это установлено И. Г. Журбенко и его сотрудниками еще в 1980-х годах [22].

Явный учет погрешностей может привести к неожиданным выводам. Так, для гамма-распределения еще Р. Фишер в 1920-х годах сравнивал по эффективности метод моментов оценки параметров и метод максимального правдоподобия, последний оказался лучше. Когда же были учтены погрешности наблюдений [1], то вывод оказался другим — в обширной области исходных данных метод моментов лучше метода максимального правдоподобия.

Большой материал по рассматриваемым вопросам представлен в весьма ценной книге [23]. Однако, по моему мнению, авторы [23] слишком много внимания уделяют нынешнему состоянию прикладной математики по сравнению с обсуждением путей развития. Кроме того, методы анализа данных предлагаются прежде всего для их

массового использования, поэтому согласно [23, гл. 2] необходимо их тщательное исследование. Однако в настоящее время бесконтрольно распространяется большое число плохо обоснованных методов (некоторые примеры ошибок даны выше). Это представляет, на мой взгляд, большую опасность, поскольку с развитием цифровизации происходит стандартизация статистического инструментария на основе стандартных пакетов прикладных статистических программ. Опасность состоит в возможности проникновения в стандартные пакеты плохо обоснованных методов. Подобные методы есть даже в лучших современных пакетах [24]. Необходимы широкие и глубокие исследования имеющихся методов анализа данных, нацеленные на создание «золотого фонда», рекомендуемого для массового использования. Пример такой попытки — система государственных стандартов по статистическим методам управления качеством продукции, прежде всего серия ГОСТов по прикладной статистике (ГОСТ 11.001-73 – ГОСТ 11.001-83). К сожалению, попытка провалилась — во многих стандартах этой системы были обнаружены грубые ошибки [25]. Причина — некомпетентность ряда разработчиков.

Очевидно, целесообразно провести анализ методов классификации, нацеленный на создание «золотого фонда». Для этого необходимо провести ряд исследований в духе описанных в статье [19]. Надо также навести порядок в терминологии: вряд ли допустимо, чтобы одна и та же область имела массу названий — кластер-анализ, распознавание образов без учителя, таксономия, автоматическая классификация и т.д.

Нужно обсуждать и показатели качества классификации. Так, например, при классификации на два класса в качестве подобного показателя часто используют долю ошибочно классифицированных объектов, что нерационально. Если доля одного из классов сравнительно мала, то вполне обоснованный алгоритм может по этому показателю оказаться хуже тривиального, согласно которому следует отнести все объекты к более многочисленному классу. Так, ряд работ группы И. М. Гельфанда посвящен прогнозированию исхода инфаркта миокарда (использовали алгоритм «Кора-3»). Если для больного прогнозировался неблагоприятный исход (смерть), то за больным следовало установить специальное наблюдение и применять интенсивное лечение — такова практическая польза применения здесь метода классификации. Ясно, что риск смерти целесообразнее несколько переоценить, чем недооценить. На это и ориентировался алгоритм группы И. М. Гельфанда. А вот по доле ошибочной классификации он оказался хуже тривиального, согласно которому предлагалось считать,

что никому из больных не угрожает смерть. Одна из возможных рекомендаций [26] — сравнивать методы классификации путем пересчета на модель линейного дискриминантного анализа, в котором классы описываются многомерными нормальными распределениями с одинаковыми ковариационными матрицами. Тогда можно оценить расстояние Махаланобиса между классами и с его помощью сравнивать методы классификации — чем это расстояние больше, тем метод классификации лучше. Пусть доля правильно классифицированных объектов в первом классе есть α , а во втором — β . Тогда оценкой расстояния Махаланобиса между классами является

$$d = \Phi^{-1}(\alpha) + \Phi^{-1}(\beta), \quad (1)$$

а в качестве оценки прогностической силы алгоритма классификации вместо доли правильных прогнозов рекомендуется использовать прогностическую силу

$$\gamma = \Phi\left(\frac{d}{2}\right) = \Phi\left(\frac{\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)}{2}\right),$$

где $\Phi(\cdot)$ — функция нормального распределения с нулевым математическим ожиданием и единичной дисперсией, а Φ^{-1} — функция, обратная к Φ .

5. В большинстве случаев анализируются данные о выборке в целях переноса на более широкую совокупность, в частности, для прогноза поведения вновь появляющегося объекта. Необходимо указывать точностные характеристики метода, т.е. точность оценивания по выборке параметров и характеристик модели. В вероятностных моделях это делается с помощью доверительных множеств, которыми обычно являются доверительные интервалы.

С прикладной точки зрения метод, для которого неизвестны точностные характеристики, является недостаточно разработанным, другими словами, поисковым, экспериментальным, эвристическим. Его нельзя рекомендовать для массового использования. Его применение может оказаться полезным, а может привести к грубым ошибкам, т.е. он является «магическим» в терминологии В. Н. Тутубалина [27].

Суть дела проста: интуиция обманывает, представляет метод гораздо более точным, чем он есть на самом деле. Современному научно-техническому уровню отвечают работы, в которых наряду с точечными оценками даны доверительные границы. Отходят от этого требования как несведущие в статистике лица, так и, к сожалению, отдельные преподаватели высшей школы, в том числе университетов, что объясняется, видимо, сочетанием «академичности» и отрыва от

массы специалистов, обрабатывающих реальные данные.

При публикации результатов статистического анализа реальных данных необходимо указывать их точность (доверительные интервалы). Иначе невозможно использование этих результатов в дальнейших исследованиях в качестве исходных данных (поскольку неизвестны «допустимые отклонения исходных данных» — [17, 18]), а также сравнение результатов различных исследований. К сожалению, данные социологических, медицинских и иных исследований часто публикуются без указания их точностных характеристик. Потом с содержательной точки зрения (т.е. с точки зрения конкретной прикладной ситуации) обсуждают, например, причины различия показателей для двух групп, в то время как статистические данные, которые можно извлечь из работы, не позволяют сделать заключение о значимости рассматриваемого различия. Имеется в виду частный случай задачи, рассмотренной выше — проверка однородности для независимых выборок из двух биномиальных распределений. Если есть две выборки объема 100, в первой положительных ответов — 47 %, а во второй — 61 %, то различие незначимо (на уровне значимости 5 %). Но социолог этого не знает — точностные характеристики не указаны — и начинает наводить теорию В журнале «Химия и жизнь» (1976, № 4, с. 112 – 113) всерьез обсуждалась связь между специальностью ученого и знаком Зодиака, под которым он родился, хотя элементарный подсчет по критерию хи-квадрат показывает, что никакой связи нет (см. подробный разбор в [28, гл. 2]). Достоинно сожаления, что отдельные специалисты по математическим методам в социологии всерьез воспринимают так называемый «детерминационный анализ» [29], в котором используются сравнительно малые по численности группы и игнорируются точностные характеристики, что приводит к получению неадекватных выводов (отметим, что с математической точки зрения «детерминационный анализ» покрывается одним из параграфов книги Г. С. Лбова [30]). Малограмотны и претенциозны высказывания о статистических методах в науковедении в книге [31]... Впрочем, все ошибки не перечислишь. Напомним хотя бы о хроническом непонимании области применимости критерия Колмогорова, разобранным в статье [32] и других работах.

По моему мнению, неточны слова К. Джини [33, с. 29]: «Нельзя предпочесть метод, который не отвечает определенной цели, методу, отвечающему цели, только на том основании, что в одном случае вычислена, а в другом еще не вычислена вероятная ошибка». Как можно знать, что «метод отвечает цели», если его точность неиз-

вестна? В частности, правильней принять решение априори, лучше ли он тривиального метода, а на данные вообще не смотреть. Из сказанного ясно, что неверно и мнение Е. С. Вентцель [34] о том, что построению доверительных интервалов не следует уделять большого внимания.

В последние десятилетия получили распространение «невероятностные методы обработки данных», или «анализ данных» (в узком смысле). Типичные публикации по анализу данных — статья [21] и книги [29, 30, 35]. Как правило, методы анализа данных — это эвристические методы, вопрос о точностных характеристиках которых даже не ставится. Справедливо сказано в [35, с. 15]: «Анализ данных применяется на первых этапах теоретического познания исследуемого явления». Очевидно, за первыми этапами должны следовать дальнейшие, имеющие целью развитие вероятностно-статистической теории, т.е. построение адекватной вероятностной модели явления и на ее основе — теоретически обоснованных правил принятия решений (например, решений о необходимости наладки технологического процесса). Таким образом, анализ данных содержит методы, которые можно сравнить с «временками»: они первыми появляются на месте будущих зданий, а после окончания строительства подлежат сносу. Это поисковые, магические, а не научно обоснованные методы, их нельзя рекомендовать для широкого использования, включать в нормативно-техническую документацию — до оценок точности получаемых с их помощью решений, что в большинстве случаев возможно лишь с помощью вероятностной модели. Последняя необходима, если полученные по выборке результаты распространяются на более широкую совокупность. Если же интересующие специалиста методы включены в исследование, то точность понимается в соответствии с теорией устойчивости [17, 18]. Реальная опасность состоит в том, что в условиях современного обилия публикаций и программ, оборотной стороной которого является относительное невежество специалистов (нельзя знать и 5 % от более чем миллиона актуальных к настоящему времени публикаций по математической статистике), распространение получают недостаточно обоснованные методы анализа данных. Временку легче построить, чем здание..., в строительстве временки стоят десятки лет. Как говорят, «нет ничего более постоянного, чем временное».

Отметим, что математические методы исследования делятся на «разведочный анализ» и «доказательную статистику». Разведочный анализ нацелен на обнаружение нового, в то время как цель доказательной статистики — строго обосновать выводы. Например, разведочный анализ дает возможность сформулировать статистиче-

скую гипотезу, а доказательная статистика позволяет ее обосновать (принять) на выбранном заранее уровне значимости.

Многие методы анализа данных основаны на максимизации какого-либо функционала. Надо подчеркнуть, что наличие оптимизации не делает метод более научным, она — средство, а не цель. В связи с обсуждением оценивания параметров гамма-распределения [1] уже приводили примеры того, что не основанные на оптимизации методы могут быть лучше оптимизационных. Польза от экстремальной формулировки основных задач прикладной статистики состоит в основном в том, что можно едиными методами изучать асимптотическое поведение решений этих задач [36], а также единообразно строить алгоритмы их решения. Наиболее естественная оптимизационная постановка задач кластер-анализа дана А. Н. Колмогоровым (см. [17]).

6. Специфические требования к методам обработки данных возникают в связи с их «стыковкой» при последовательном выполнении [13, 19]: результаты работы предыдущего алгоритма должны удовлетворять условиям, наложенным на исходные данные последующего. Так, «восстановление пропущенных данных» по какому-либо алгоритму приводит к тому, что полученная матрица «объект-признак» не может рассматриваться как составленная из независимых случайных векторов, т.е. классическое предположение математической статистики — «наблюдения есть выборка» (конечная последовательность независимых одинаково распределенных случайных элементов) — не выполнено, следовательно, применение основанных на этом предположении методов не является обоснованным. Аналогичная ситуация возникает при «преобразовании данных», если параметры преобразования определяются по исходным данным. Неясной остается на настоящий момент обоснованность регрессионного анализа, если степень полинома, описывающего линию регрессии, подбирается по экспериментальным данным, поскольку распространенные оценки этой степени несостоятельны [37]. Продолжать можно долго. К сожалению, нельзя априори надеяться, что влияние указанных нарушений исходных предпосылок мало. Так, в критериях согласия Колмогорова, омега-квадрат и др. возникает желание вместо неизвестных параметров подставить их оценки. Этот прием аналогичен рассмотренным выше, но, в отличие от них, последствия его применения хорошо изучены. Влияние велико и не уменьшается с ростом выборки, например, при применении критерия Колмогорова для проверки нормальности процентные точки должны быть уменьшены примерно в 1,5 раза по сравнению с классическими [32].

Распространена рекомендация — разбить совокупность на однородные классы и затем анализировать каждый класс отдельно. Рекомендация рациональна [23]. Так, при обработке данных о течении острой пневмонии [38] коэффициент корреляции между возрастом и длительностью заболевания оказался сравнительно малым ($r = 0,21$). Когда же выделили группы курящих и некурящих, то в первой из них связь оказалась гораздо более выраженной ($r = 0,53$), во второй же — незначимой.

В рассмотренной задаче классы выделены по априорным соображениям. Если же дискриминирующая поверхность (разделяющая классы) строится на основе анализа экспериментальных данных, то попавшие в один класс наблюдения, вообще говоря, не образуют выборку (нарушается независимость), а распределения их не являются нормальными. Для естественной модели показано [19], что при росте объема выборки независимость в определенном смысле восстанавливается, в то время как распределение элементов кластера отнюдь не приближается к распределению соответствующего члена в смеси, описывающей исходную совокупность (в частности, плотность этого распределения равна нулю для обширной области пространства). Следовательно, нельзя применять регрессионный анализ, основанный на предположении нормальности.

7. Требования к представлению результатов статистического анализа частично рассмотрены выше. Результаты должны приводиться вместе с точностными характеристиками, с указанием конкретного метода, с помощью которого они получены, и степени его обоснованности. При использовании информационно-коммуникационных технологий следует указывать тип (марку, название) компьютера, язык программирования, время счета и другие необходимые характеристики.

Кроме указанных выше, можно сформулировать ряд иных требований к методам обработки данных и представлению их результатов [39].

О границах применимости вероятностно-статистических методов

Этой теме посвящены многочисленные публикации [8, 9, 27, 40 – 43], коснулись ее и работы автора [17, 44, 45]. Весьма кратко отмечу только два обстоятельства, не претендуя на завершенность.

1. По моему мнению, применение вероятностных методов не имеет принципиальных отличий от применений других областей математики как более старых (геометрия, дифференциальные уравнения), так и более новых (теория нечеткости [45]). Схема применения однотипна:

строят модель на основе соответствующей области математики, тем или иным способом ее обосновывают, на основе модели реального явления изучают интересующие специалистов вопросы, полученные выводы интерпретируют и используют для принятия решений. Поразительно, что отдельные авторы полностью игнорируют многочисленные способы проверки адекватности вероятностной модели.

2. Не менее поразительно, что возможность применения вероятностных моделей связывают с «темными понятиями» устойчивости частот, статистической однородности, статистического ансамбля [8, 27, 43]. Вот уже более 80 лет теория вероятностей является аксиоматической наукой — мы основываемся на аксиоматике А. Н. Колмогорова [46] (его основополагающая монография впервые издана в 1933 г. на немецком языке и в 1936 г. — на русском). В ней нет места перечисленным «темным понятиям», как и бессмысленному, вообще говоря, понятию «генеральная совокупность» (оно имеет смысл лишь в случае выбора из конечного множества). Понятие статистического ансамбля, как и выражение «теория вероятностей изучает закономерности массовых явлений» — это реликты начала XX века, когда не отделяли математическую теорию вероятностей от ее приложений. Попытки применить эти понятия сводятся к бездоказательным общим рассуждениям (другими словами, демагогии), поскольку любая научно обоснованная проверка должна опираться на вероятностную модель явления. На наш взгляд, движение одной-единственной частицы или развитие уникальной экономической системы вполне можно описывать случайными процессами (в терминологии теории вероятностей), если соответствующие вероятностные модели обоснованы. Например, как проверить, что $w(t)$ — траектория винеровского процесса? Возьмем разности $w(\tau) - w(0)$, $w(2\tau) - w(\tau)$, $w(3\tau) - w(2\tau)$, ..., где τ достаточно мало по сравнению с интервалом наблюдения. Тогда, как известно, все эти разности независимы и имеют одинаковое нормальное распределение с нулевым математическим ожиданием и дисперсией τ — при справедливости гипотезы о винеровости. Остается проверить, справедливо ли утверждение, сформулированное в предыдущей фразе, с помощью широко известных статистических критериев [1, 6, 7, 11 и др.].

О некоторых постановках задач классификации

1. Если классы полностью описаны или заданы обучающими выборками, классификацию можно рассматривать как измерение. При статистическом контроле качества единицы продук-

ции классифицируются на годные и бракованные. Врач ставит диагноз больному, относя тем самым его заболевание к одной из нозологических форм. Измерение в номинальной шкале есть разбиение объектов на классы, а в порядковой — на упорядоченные классы [17]. Ясно, что результат измерения должен быть воспроизводимым, допускать сравнение с результатами других измерений. Вообще, классификация как средство измерения должна удовлетворять требованиям, устанавливаемым метрологией. Необходимым условием этого является стандартизация правил классификации (однако это условие не является достаточным — сплошь и рядом контролирующие органы обнаруживают, что пропущенные службами контроля качества изделия не удовлетворяют требованиям соответствующих нормативных документов). Ясно, что без стандартизации правил классификации не могут работать различные автоматизированные системы управления, действующие на предприятиях и в регионах. В статистике говорят о точном определении используемых понятий, рассматриваемых совокупностей [33, 47].

Хотя с необходимостью применения стандартных классификаций обычно никто не спорит, на практике стандартизация не всегда осуществляется. Сотрудники вузов хорошо это знают, сравнивая оценки в школьных аттестатах и на экзаменах. Мне уже приходилось упоминать [45] о двух группах медиков, по определению одной из которых «затяжное течение острой пневмонии» имело место в 6 % случаев, а по мнению другой — в 60 % (для той же совокупности из 461 больного)! Неточности классификаций приводят к тому, что экономико-статистические данные имеют относительные ошибки 5 – 10 % [47].

2. В ряде случаев «мы хотим разбить объекты на группы независимо от того, естественны границы разбиения или нет» [48, с. 437]. Типичные примеры — использования интервалов группировки в статистике, разбиение студентов специальности по учебным группам.

3. «Проблема классификации (в узком смысле слова — А. О.) состоит в выяснении по эмпирическим данным, насколько элементы «группируются» или распадаются на изолированные «скопления», «кластеры»» [48, с. 467]. Рассматриваемую область прикладной статистики естественно называть кластер-анализом. В этой области наиболее обоснованными являются вероятностно-статистические методы, известные как методы расщепления смесей [19]. При использовании тех или иных алгоритмов возникает проблема «реальности кластера» [19]. Дело в том, что алгоритм кластер-анализа можно применить к любым исходным данным, в том числе к выборке из однородной совокупности. В последнем слу-

чае, очевидно, результат работы алгоритма не будет иметь реального смысла. Как отличить эту ситуацию от противоположной, когда совокупность действительно разбивается на кластеры? Приведем пример ошибочного применения кластер-анализа.

Качество одного из продуктов нефтехимии — фенола — характеризуют 13 показателей. На их измерения тратятся большие средства. Идея состоит в том, чтобы разбить признаки на группы и из каждой группы оставить только один, при этом «каждый из признаков внутри одной группы говорит об образцах почти одно и то же» [49, с. 23]. Последнее означает, что коэффициенты корреляции между признаками одной группы близки к единице. По экспериментальным данным нашли матрицу выборочных коэффициентов корреляции [49, с. 25]. Максимальный по величине коэффициент корреляции равен 0,85, следующий за ним — 0,46. Отсюда ясно, что только два признака из 13 связаны между собой настолько, что имеет смысл прогнозировать значение одного из них по другому, да и для них прогнозирование не слишком хорошее. Однако это не смущает Ю. П. Адлера, который, не колеблясь, применяет метод корреляционных плеяд и получает шесть групп. Одна из них состоит из двух показателей, коэффициент корреляции между которыми равен 0,21 [49, с. 25], т.е. с помощью одного из них можно объяснить лишь 4 % дисперсии второго. Обоснованный (с позиций прикладной статистики) ответ в рассматриваемой задаче таков: показатели практически нельзя объединить в группы (за исключением двух, коэффициент корреляции между которыми равен 0,85); чтобы не потерять информацию, надо измерять не менее 12 показателей. Однако Ю. П. Адлер считает, что достаточно шести — по одному из группы [49, с. 24]. Это — введение заказчика в заблуждение с использованием авторитета математических методов. Интересно подсчитать убытки, вызванные такой рекомендацией.

Если кластеры являются реальными, то любой разумный алгоритм кластер-анализа их достаточно точно выделит. Другими словами, результат кластер-анализа должен быть устойчив относительно выбора алгоритма [17, 18]. Следовательно, для выделения реальных кластеров можно рекомендовать наиболее простой в определенном смысле алгоритм, например, требующий наименьших вычислений, скажем, алгоритм ближнего соседа [1]. Затем следует проверить устойчивость полученных кластеров по отношению к допустимым отклонениям исходных данных [19].

Приведем пример. Мы обрабатывали анкеты (типа социологических) способных к математике школьников [50]. Для кластер-анализа

признаков, измеренных в номинальных шкалах, был выбран алгоритм [21], который мы сочли под влиянием [21] наиболее перспективным и обоснованным. Реализация алгоритма на компьютере и счет заняли около полугода. Позже я за полтора часа обработал вручную те же данные по упомянутому выше алгоритму ближнего соседа. Результаты (дендрограммы) практически совпали. Более того, алгоритм ближнего соседа дал дополнительную информацию о структуре данных. Итак, цена ошибочного выбора алгоритма — полгода лишней работы плюс стоимость машинного времени (вторая составляющая в рассматриваемое время была заметной).

Самый радикальный способ сократить затраты на кластер-анализ — заранее объявить совокупность однородной. Так, Ю. Н. Тюрин [51] пишет: «При проведении экспертного опроса обычно считают, что по интересующему предмету существует истинная точка зрения». Если же выявились кластеры различных мнений, то «надо признать, что экспертный опрос не достиг окончательной цели» [51, с. 11]. По моему мнению, это слишком категоричное заявление. Оно может повлечь исключение из процедур обработки экспертных данных этапа кластер-анализа, а следовательно, привести к ошибкам в содержательных областях. На практике мнения экспертов зачастую разделяются (например, мнения научных работников и производственников). Мы полагаем, что при применении экспертных технологий необходим этап классификации мнений экспертов (отметим, что [52] в модели люсианов порождения экспертных оценок удалось из вероятностно-статистических соображений указать ограничение сверху на диаметр кластера, т.е. обосновать выбор итогового разбиения из дендрограммы).

Один из разделов статьи Н. Бурбаки «Архитектура математики», основополагающей для многотомной серии «Элементы математики», называется «Стандартизация математических орудий» [53, с. 253]. Наша задача — стандартизовать такое мощное орудие, как методы классификации.

В настоящей статье раскрыт ряд положений колонки редколлегии [54].

Благодарности

Автор выражает благодарность Д. С. Шмерлингу за конструктивную критику.

ЛИТЕРАТУРА

1. Орлов А. И. Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
2. Орлов А. И. Характеризация средних величин шкалами измерения / Научный журнал КубГАУ. 2017. № 134. С. 877 – 907.

3. Енюков И. С. Методы оцифровки не количественных признаков. В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. — М.: Наука, 1980. С. 309 – 316.
4. Александров В. В., Горский Н. Д. Алгоритмы и программы структурного метода обработки данных. — Л.: Наука, 1983. — 208 с.
5. Саати Т. Л. Принятие решений. Метод анализа иерархий. — М.: Радио и связь, 1989. — 316 с.
6. Гаек Я., Шидак З. Теория ранговых критериев / Пер. с англ. — М.: Наука, 1971. — 376 с.
7. Холлендер М., Вульф Д. Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
8. Алимов Ю. И. Альтернатива методу математической статистики. — М.: Знание, 1980. — 64 с.
9. Малиновский Л. Г. Анализ статистических связей: модельно-конструктивный подход. — М.: Наука, 2002. — 687 с.
10. Орлов А. И. О методах проверки однородности двух независимых выборок / Заводская лаборатория. Диагностика материалов. 2020. Т. 86. № 3. С. 67 – 76. DOI: 10.26896/1028-6861-2020-86-3-67-76.
11. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
12. Орлов А. И. Структура непараметрической статистики (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 7. С. 62 – 72.
13. Загоруйко Н. Г., Орлов А. И. Некоторые нерешенные математические задачи прикладной статистики. В сб.: Современные проблемы кибернетики (прикладная статистика). — М.: Знание, 1981. С. 53 – 63.
14. Орлов А. И., Миронова Н. Г., Фомин В. Н., Черномордик О. М. Методика. Проверка однородности двух выборок параметров продукции при оценке ее технического уровня и качества. — М.: ВНИИСтандартизации, 1987. — 116 с.
15. Орлов А. И. Реальные и номинальные уровни значимости при проверке статистических гипотез / Научный журнал КубГАУ. 2015. № 114. С. 42 – 54.
16. Орлов А. И. Состоятельные критерии проверки абсолютной однородности независимых выборок / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 11. С. 66 – 70.
17. Орлов А. И. Устойчивость в социально-экономических моделях. — М.: Наука, 1979. — 296 с.
18. Орлов А. И. Устойчивые экономико-математические методы и модели. Разработка и развитие устойчивых экономико-математических методов и моделей для модернизации управления предприятиями. — Saarbrücken (Germany), LAP (Lambert Academic Publishing), 2011. — 436 с.
19. Орлов А. И. Некоторые вероятностные вопросы теории классификации. В сб.: Прикладная статистика. — М.: Наука, 1983. С. 166 – 179.
20. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1986. — 288 с.
21. Куперштох В. Л., Миркин Б. Г., Трофимов В. А. Сумма внутренних связей как показатель качества классификации / Автоматика и телемеханика. 1976. № 3. С. 91 – 98.
22. Орлов А. И. Метод статистических испытаний в прикладной статистике / Заводская лаборатория. Диагностика материалов. 2019. Т. 85. № 5. С. 67 – 79. DOI: 10.26896/1028-6861-2019-85-5-67-79.
23. Блехман И. И., Мышкис А. Д., Пановко Я. Г. Механика и прикладная математика: логика и особенности приложенной математики. Изд. 2-е, испр. и доп. — М.: Наука, 1990. — 360 с.
24. Орлов А. И. Статистические пакеты — инструменты исследователя / Заводская лаборатория. Диагностика материалов. 2008. Т. 74. № 5. С. 76 – 78.
25. Орлов А. И. Сертификация и статистические методы (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 1997. Т. 63. № 3. С. 55 – 62.
26. Орлов А. И. Прогностическая сила — наилучший показатель качества алгоритма диагностики / Научный журнал КубГАУ. 2014. № 99. С. 33 – 49.

27. **Тутубалин В. Н.** Теория вероятностей в естествознании. — М.: Знание, 1972. — 64 с.
28. **Орлов А. И.** Эконометрика. Изд. 4-е, доп. и перераб. — Ростов-на-Дону: Феникс, 2009. — 572 с.
29. **Чесноков С. В.** Детерминационный анализ социально-экономических данных. Изд. 2-е, испр. и доп. — М.: URSS, 2009. — 168 с.
30. **Лбов Г. С.** Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981. — 160 с.
31. **Хайтун С. Д.** Наукометрия: Состояние и перспективы. — М.: Наука, 1983. — 344 с.
32. **Орлов А. И.** Непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат и ошибки при их применении / Научный журнал КубГАУ. 2014. № 97. С. 32 – 45.
33. **Джини К.** Логика в статистике. — М.: Статистика, 1973. — 128 с.
34. **Вентцель Е. С.** Методологические особенности прикладной математики на современном этапе. В сб.: Математики о математике. — М.: Знание, 1982. С. 37 – 55.
35. **Миркин Б. Г.** Анализ качественных признаков и структур. — М.: Статистика, 1980. — 319 с.
36. **Орлов А. И.** Предельная теория решений экстремальных статистических задач / Научный журнал КубГАУ. 2017. № 133. С. 579 – 600.
37. **Орлов А. И.** Оценка размерности модели в регрессии. В сб.: Алгоритмическое и программное обеспечение прикладного статистического анализа. — М.: Наука, 1980. С. 92 – 99.
38. **Рабухин А. Е., Сильвестров В. П., Орлов А. И. и др.** Результаты лечения больных острой пневмонией. В сб.: Актуальные вопросы клинической и экспериментальной медицины. — М.: 4 ГУ МЗ СССР, 1978. С. 132 – 138.
39. **Орлов А. И., Миронова Н. Г., Фомин В. Н., Черчинцев А. Н.** Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. — М.: ВНИИСтандартизации, 1987. — 62 с.
40. **Купцов В. И.** Детерминизм и вероятность. — М.: Политиздат, 1976. — 256 с.
41. **Сачков Ю. В.** Вероятностная революция в науке (Вероятность, случайность, независимость, иерархия). — М.: Научный мир, 1999. — 144 с.
42. **Сачков Ю. В.** Введение в вероятностный мир. — М.: Наука, 1971. — 208 с.
43. **Тутубалин В. Н.** Границы применимости (вероятностно-статистические методы и их возможности). — М.: Знание, 1977. — 64 с.
44. **Орлов А. И.** О развитии прикладной статистики. В сб.: Современные проблемы кибернетики (прикладная статистика). — М.: Знание, 1981. С. 3 – 14.
45. **Орлов А. И.** Математика нечеткости / Наука и жизнь. 1982. № 7. С. 60 – 67.
46. **Колмогоров А. Н.** Основные понятия теории вероятностей. Изд. 2-е. — М.: Наука, 1974. — 120 с.
47. **Моргенштерн О.** О точности экономико-статистических наблюдений. — М.: Статистика, 1968. — 293 с.
48. **Кендалл М. Дж., Стьюарт А.** Многомерный статистический анализ и временные ряды. — М.: Наука, 1976. — 736 с.
49. **Адлер Ю. П.** Управление качеством: статистический подход. — М.: Знание, 1979. — 51 с.
50. **Орлов А. И., Гусейнов Г. А.** Математические методы в изучении способных к математике школьников. В сб.: Исследования по вероятностно-статистическому моделированию реальных систем. — М.: ЦЭМИ АН СССР, 1977. С. 80 – 93.
51. **Тюрин Ю. Н.** О математических задачах в экспертных оценках / Экспертные оценки. Вопросы кибернетики. Вып. 58. — М.: Научный совет АН СССР по комплексной проблеме «Кибернетика», 1979. С. 7 – 16.
52. **Орлов А. И.** Теория лосианов / Научный журнал КубГАУ. 2014. № 101. С. 275 – 304.
53. **Бурбаки Н.** Очерки по истории математики. — М.: ИЛ, 1963. — 292 с.
54. **Орлов А. И.** Роль методологии в математических методах исследования / Заводская лаборатория. Диагностика материалов. 2019. Т. 85. № 7. С. 5 – 6.

REFERENCES

1. **Orlov A. I.** Applied Statistics. — Moscow: Ékzamen, 2006. — 671 p. [in Russian].
2. **Orlov A. I.** Characterization of average values by means of measurement scales / Nauch. Zh. KubGAU. 2017. N 134. P 877 – 907 [in Russian].
3. **Enyukov I. S.** Methods for digitizing non-quantitative attributes / Algorithmic and software for applied statistical analysis. — Moscow: Nauka, 1980. P. 309 – 316 [in Russian].
4. **Aleksandrov V. V., Gorskii N. D.** Algorithms and programs for the structural method of data processing. — Leningrad: Nauka, 1983. — 208 p. [in Russian].
5. **Saati T. L.** Decisionmaking. Hierarchy Analysis Method. — Moscow: Radio i svyaz', 1993. — 278 p. [in Russian].
6. **Hajek Ja., Sidak Zb.** Theory of rank tests. — Prague: Academia. Publishing House of the Czechoslovak Academy of Sciences, 1967. — 376 p.
7. **Hollander M., Wolfe D. A.** Nonparametric statistical methods. — New York – London – Sydney – Toronto: John Wiley and Sons, 1973. — 503 p.
8. **Alimov Yu. I.** Alternative to the method of mathematical statistics. — Moscow: Znanie, 1980. — 64 p. [in Russian].
9. **Malinovskii L. G.** Analysis of statistical relationships: a model-constructive approach. — Moscow: Nauka, 2002. — 687 p. [in Russian].
10. **Orlov A. I.** On the methods of testing the homogeneity of two independent samples / Zavod. Lab. Diagn. Mater. 2020. Vol. 86. N 3. P. 67 – 76. DOI: 1026896/1028-6861-2020-86-3-67-76 [in Russian].
11. **Bol'shev L. N., Smirnov N. V.** Tables of mathematical statistics. 3rd edition. — Moscow: Nauka, 1983. — 416 p. [in Russian].
12. **Orlov A. I.** Structure of Nonparametric Statistics (generalizing paper) / Zavod. Lab. Diagn. Mater. 2015. Vol. 81. N 7. P. 62 – 72 [in Russian].
13. **Zagoruiko N. G., Orlov A. I.** Some unsolved mathematical problems of applied statistics / Modern problems of cybernetics (applied statistics). — Moscow: Znanie, 1981. P. 53 – 63 [in Russian].
14. **Orlov A. I., Mironova N. G., Fomin V. N., Chernomordik O. M.** Methodology. Testing of the homogeneity of two samples of product parameters in estimation its technical level and quality. — Moscow: VNIISTandartizatsii, 1987. — 116 p. [in Russian].
15. **Orlov A. I.** Real and nominal significance levels when testing statistical hypotheses / Nauch. Zh. KubGAU. 2015. N 114. P. 42 – 54 [in Russian].
16. **Orlov A. I.** Consistent Tests of Absolute Homogeneity of Independent Samplings / Zavod. Lab. Diagn. Mater. 2012. Vol. 78. N 11. P. 66 – 70 [in Russian].
17. **Orlov A. I.** Stability in socio-economic models. — Moscow: Nauka, 1979. — 296 p. [in Russian].
18. **Orlov A. I.** Stable economic and mathematical methods and models. Design and development of stable economic and mathematical methods and models for the modernization of enterprise management. — Saarbrücken (Germany), LAP (Lambert Academic Publishing), 2011. — 436 p. [in Russian].
19. **Orlov A. I.** Some probabilistic questions of classification theory / Applied statistics. — Moscow: Nauka, 1983. P. 166 – 179 [in Russian].
20. **Tikhonov A. N., Arsenin V. Ya.** Methods for solving incorrect problems. — Moscow: Nauka, 1986. — 288 p. [in Russian].
21. **Kupershtokh V. L., Mirkin B. G., Trofimov V. A.** The sum of internal relationships as an indicator of the quality of classification / Avtomat. Telemekh. 1976. N 3. P. 91 – 98 [in Russian].
22. **Orlov A. I.** Statistical simulations method in applied statistics / Zavod. Lab. Diagn. Mater. 2019. Vol. 85. N 5. P. 67 – 79 [in Russian].

23. **Blekhman I. I., Myshkis A. D., Panovko Ya. G.** Mechanics and Applied Mathematics: Logic and Features of Mathematics Applications. 2nd edition. — Moscow: Nauka, 1990. — 360 p. [in Russian].
24. **Orlov A. I.** Statistic Packages: Tools of the Researcher / *Zavod. Lab. Diagn. Mater.* 2008. Vol. 74. N 5. P. 76 – 78 [in Russian].
25. **Orlov A. I.** Certification and statistical methods (generalizing article) / *Zavod. Lab. Diagn. Mater.* 1997. Vol. 63. N 3. P. 55 – 62 [in Russian].
26. **Orlov A. I.** Predictive power — the best indicator of the quality of the diagnostic algorithm / *Nauch. Zh. KubGAU.* 2014. N 99. P. 33 – 49 [in Russian].
27. **Tutubalin V. N.** Probability Theory in Natural Science. — Moscow: Znanie, 1972. — 64 p. [in Russian].
28. **Orlov A. I.** Econometrics. 4th edition. — Rostov-on-Don: Feniks, 2009. — 572 p. [in Russian].
29. **Chesnokov S. V.** Determinative analysis of socio-economic data. 2nd edition. — Moscow: URSS, 2009. — 168 p. [in Russian].
30. **Lbov G. S.** Methods for processing heterogeneous experimental data. — Novosibirsk: Nauka, 1981. 160 p. [in Russian].
31. **Khaitun S. D.** Scientometrics: Status and Prospects. — Moscow: Nauka, 1983. — 344 p. [in Russian].
32. **Orlov A. I.** Nonparametric goodness-of-fit Kolmogorov, Smirnov, omega-square tests and the errors in their application / *Nauch. Zh. KubGAU.* 2014. N 97. P. 32 – 45 [in Russian].
33. **Gini C.** Logic in statistics. — Moscow: Statistika, 1973. — 128 p. [Russian translation].
34. **Ventsel' E. S.** Methodological features of applied mathematics at the present stage / *Mathematicians about mathematics.* — Moscow: Znanie, 1982. P. 37 – 55 [in Russian].
35. **Mirkin B. G.** Analysis of quality attributes and structures. — Moscow: Statistika, 1980. — 319 p. [in Russian].
36. **Orlov A. I.** The limit theory of the solutions of extremal statistical problems / *Nauch. Zh. KubGAU.* 2017. N 133. P. 579 – 600 [in Russian].
37. **Orlov A. I.** Estimation of the dimensionality of a model in regression / *Algorithmic and software for applied statistical analysis.* — Moscow: Nauka, 1980. P. 92 – 99 [in Russian].
38. **Rabukhin A. E., Sil'vestrov V. P., Orlov A. I., et al.** The results of treatment of patients with acute pneumonia / *Current issues of clinical and experimental medicine.* — Moscow: 4 GU MZ SSSR, 1978. P. 132 – 138 [in Russian].
39. **Orlov A. I., Mironova N. G., Fomin V. N., Cherchintsev A. N.** Recommendations. Applied statistics. Data processing methods. Basic requirements and characteristics. — Moscow: VNIISstandartizatsii, 1987. — 62 p. [in Russian].
40. **Kuptsov V. I.** Determinism and probability. — Moscow: Politizdat, 1976. — 256 p. [in Russian].
41. **Sachkov Yu. V.** Probabilistic revolution in science (Probability, chance, independence, hierarchy). — Moscow: Nauchnyi mir, 1999. — 144 p. [in Russian].
42. **Sachkov Yu. V.** Introduction to the Probabilistic World. — Moscow: Nauka, 1971. — 208 p. [in Russian].
43. **Tutubalin V. N.** Boundaries of applicability (probabilistic-statistical methods and their capabilities). — Moscow: Znanie, 1977. — 64 p. [in Russian].
44. **Orlov A. I.** On the development of applied statistics / *Modern problems of cybernetics (applied statistics).* — Moscow: Znanie, 1981. P. 3 – 14 [in Russian].
45. **Orlov A. I.** Mathematics of fuzziness / *Nauka i zhizn'.* 1982. N 7. P. 60 – 67 [in Russian].
46. **Kolmogorov A. N.** Basic concepts of probability theory. 2nd edition. — Moscow: Nauka, 1974. — 120 p. [in Russian].
47. **Morgenshtern O.** On the accuracy of economic and statistical observations. — Moscow: Statistika, 1968. — 293 p. [in Russian].
48. **Kendall M. G., Stuart A.** Multivariate Statistical Analysis and Time Series. — Moscow: Nauka, 1976. — 736 p. [in Russian].
49. **Adler Yu. P.** Quality Management: A Statistical Approach. — Moscow: Znanie, 1979. — 51 p. [in Russian].
50. **Orlov A. I., Guseinov G. A.** Mathematical methods in the study of mathematics-capable schoolchildrens / *Research on probabilistic and statistical modeling of real systems.* — Moscow: TsÉMI AN SSSR, 1977. P. 80 – 93 [in Russian].
51. **Tyurin Yu. N.** About mathematical problems in expert estimators / *Expert assessments. Cybernetics. Issue 58.* — Moscow: Nauchnyi soviet AN SSSR po kompleksnoi probleme “Kibernetika”, 1979. P. 7 – 16 [in Russian].
52. **Orlov A. I.** Theory of lusians / *Nauch. Zh. KubGAU.* 2014. N 101. P. 275 – 304 [in Russian].
53. **Burbaki N.** Essays on the History of Mathematics. — Moscow: IL, 1963. — 292 p. [Russian translation].
54. **Orlov A. I.** The role of methodology in mathematical methods of research / *Zavod. Lab. Diagn. Mater.* 2019. Vol. 85. N 7. P. 5 – 6 [in Russian].