СОВРЕМЕННАЯ ПРИКЛАДНАЯ СТАТИСТИКА

Дается критический анализ современного состояния прикладной статистики. Обсуждаются тенденции развития статистических методов.

**1. Что дает прикладная статистика народному хозяйству?**

Так называлась статья [1], в которой приводились многочисленные примеры успешного использования методов прикладной математической статистики. при решении практических задач. Перечень примеров можно продолжать практически безгранично. Из работ 1997 г. достаточно сослаться на обобщающую статью В.Г.Горского [2], на диссертацию А.Н.Гуды [3]. По данным Института информации Гарфилда (США) каждая из основополагающих книг В.В.Налимова [4,5] цитировалась не менее 1000 раз (см. также монографию [6, с.270,274,373]). В любом номере журнала "Заводская лаборатория" есть работы, в которых те или иные методы прикладной статистики применяются для решения прикладных задач. Не раз публиковались в нашем журнале и обобщающие статьи [7-9].

Поэтому бесспорно совершенно, что методы прикладной статистики успешно применяются в различных отраслях народного хозяйства, практически во всех областях науки. Согласно докладу [10, с.157-158] в 1988 г. затраты в СССР на статистический анализ данных оценивались в 2 миллиарда рублей ежегодно.

Большая практическая значимость прикладной статистики оправдывает целесообразность проведения работ по ее методологии, в которых эта область научной и прикладной деятельности рассматривалась бы как целое, "с высоты птичьего полета". Чтобы иметь возможность обсуждения тенденций развития статистических методов, кратко рассмотрим их историю.

**2. Об истории прикладной статистики**

Типовые примеры раннего этапа применения статистических методов описаны в Ветхом Завете (см., например, Книгу Чисел). С математической точки зрения они сводились к подсчетам числа попаданий значений наблюдаемых признаков в определенные градации. В дальнейшем результаты стали представлять в виде таблиц и диаграмм, как это и сейчас делает Госкомстат РФ. Надо признать, что по сравнению с Ветхим Заветом есть прогресс - в Библии не было таблиц. Однако нет продвижения по сравнению с работами российских статистиков конца девятнадцатого - начала двадцатого века (типовой монографией тех времен можно считать книгу [11], которая в настоящее время еще легко доступна).

Сразу после возникновения теории вероятностей (Паскаль, Ферма, 17 век) вероятностные модели стали использоваться при обработке статистических данных. Например, изучалась частота рождения мальчиков и девочек, было установлено отличие вероятности рождения мальчика от 0.5, анализировались причины того, что в парижских приютах эта вероятность не та, что в самом Париже, и т.д. Имеется достаточно много публикаций по истории теории вероятностей, однако в некоторых из них имеются неточные утверждения, что заставило академика Украинской АН Б.В.Гнеденко включить в последнее издание своего курса [12] главу по истории математики случайного.

В 1794 г. (по другим данным - в 1795 г.) К.Гаусс разработал метод наименьших квадратов, один из наиболее популярных ныне статистических методов, и применил его при расчете орбиты астероида Церера - для борьбы с ошибками астрономических наблюдений [13]. В 19 веке заметный вклад в развитие практической статистики внес бельгиец Кетле, на основе анализа большого числа реальных данных показавший устойчивость относительных статистических показателей, таких, как доля самоубийств среди всех смертей [14]. Интересно, что основные идеи статистического приемочного контроля и сертификации продукции обсуждались академиком Буняковским и применялись в российской армии еще в середине 19 в.[15]. Статистические методы управления качеством, сертификации и классификации продукции сейчас весьма актуальны [16].

Современный этап развития прикладной статистики можно отсчитывать с 1900 г., когда англичанин К.Пирсон основан журнал "Biometrika". Первая треть ХХ в. прошла под знаком параметрической статистики. Изучались методы, основанные на анализе данных из параметрических семейств распределений, описываемых кривыми семейства Пирсона. Наиболее популярным было нормальное (гауссово) распределение. Для проверки гипотез использовались критерии Пирсона, Стьюдента, Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи планирования эксперимента.

Разработанную в первой трети ХХ в. теорию будем называть параметрической статистикой, поскольку ее основной объект изучения - это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами. Как правило, нельзя указать каких-либо веских причин, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением, и т.д. Однако в подавляющем большинстве реальных ситуаций подобных моделей нет, и приближение реального распределения с помощью кривых из семейства Пирсона или его подсемейств - чисто формальная операция.

Именно из таких соображений критиковал параметрическую статистику академик С.Н.Бернштейн в 1927 г. в своем докладе на Всероссийском съезде математиков [17]. Однако эта теория, к сожалению, до сих пор остается основой преподавания статистических методов и продолжает использоваться основной массой прикладников, остающихся далекими от новых веяний в статистике. Почему так происходит? Чтобы попытаться ответить на этот вопрос, обратимся к наукометрии.

**3. Наукометрия прикладной статистики**

Проведенный несколько лет назад анализ прикладной статистики как области научно-практической деятельности показал, в частности, что актуальными для специалистов в настоящее время являются не менее чем 100 тысяч публикаций (подробнее см. статьи [9,18,19]). Реально же каждый из нас знаком с существенно меньшим количеством книг и статей. Так, в известном трехтомнике Кендалла и Стьюарта [20-22] всего около 2 тысяч литературных ссылок. При всей очевидности соображений о многократном дублировании в публикациях ценных идей приходится признать, что каждый специалист по прикладной статистике владеет лишь небольшой частью накопленных в этой области знаний. Не удивительно, что приходится постоянно сталкиваться с игнорированием или повторением ранее полученных результатов, с уходом в тупиковые (с точки зрения практики) направления исследований, с беспомощностью при обращении к реальным данным, и т.д. Все это - одно из проявлений адаптационного механизма торможения развития науки, о котором еще 30 лет назад писали В.В.Налимов и другие науковеды (см., например, [23]).

Традиционный предрассудок состоит в том, что каждый новый результат, полученный исследователем - это кирпич в непрерывно растущее здание науки, который непременно будет проанализирован и использован научным сообществом. Реальная ситуация - совсем иная. Основа профессиональных знаний исследователя и инженера закладывается в период обучения. Затем они пополняются в том узком направлении, в котором работает специалист. Следующий этап - их тиражирование новому поколению. В результате вузовские учебники отстоят от современного развития на десятки лет. Так, учебники по математической статистике, по нашей экспертной оценке, в основном соответствуют 40-60-м годам ХХ в. А потому тем же годам соответствует большинство вновь публикуемых исследований и тем более - прикладных работ. Одновременно приходится признать, что результаты, не вошедшие в учебники, независимо от их ценности почти все забываются.

Активно продолжается развитие тупиковых направлений. Психологически это понятно. Приведу пример из своего опыта. В свое время по заказу Госстандарта я разработал методы оценки параметров гамма-распределения [24]. Поэтому мне близки и интересны работы по оцениванию параметров по выборкам из распределений, принадлежащих тем или иным параметрическим семействам, понятия функции максимального правдоподобия, эффективности оценок, использование неравенства Рао-Крамера и т.д. К сожалению, я знаю, что это - тупиковая ветвь, поскольку реальные данные не подчиняются каким-либо параметрическим семействам, надо применять иные статистические методы, о которых речь пойдет ниже. Понятно, что специалистам по параметрической статистике, потратившим многие годы на совершенствование в своей области, психологически трудно согласиться с этим утверждением. В том числе и мне.

**4. Точки роста**

Отечественная литература по прикладной статистике столь же необозрима, как и мировая. Только в секции "Математические методы исследования" журнала "Заводская лаборатория" с 1960-х годов опубликовано более 1000 статей. Не будем даже пытаться перечислять коллективы исследователей или основные монографии в этой области (впрочем, см. статью [19]). Отметим только два издания. По моему мнению, наилучшей отечественной книгой по прикладной статистике является сборник статистических таблиц Л.Н.Большева и Н.В.Смирнова [25] с подробными комментариями, играющими роль учебника и справочника. В распространенном трехтомном справочном издании [26-28] под редакцией С.А.Айвазяна содержится полезная информация о многих направлениях прикладной статистики. Однако в изложении имеется сравнительно много погрешностей, поэтому пользоваться книгами [26-28] приходится с осторожностью.

Основная цель настоящей статьи - выделить и обсудить "точки роста" прикладной статистики, те ее направления, которые представляются перспективными в будущем, но пока отодвинуты на задний план традиционными постановками.

Несколько лет назад при описании современного этапа развития статистических методов нами были выделены [29] пять актуальных направлений, в которых развивается современная прикладная статистика, т.е. пять "точек роста": непараметрика, робастность, бутстреп, интервальная статистика, статистика объектов нечисловой природы. Обсудим их.

**5. Непараметрическая статистика**

В первой трети ХХ в., одновременно с параметрической статистикой, в работах Спирмена и Кендалла появились первые непараметрические методы, основанные на коэффициентах ранговой корреляции, носящих ныне имена этих статистиков. Но непараметрика, не делающая нереалистических предположений о том, что функции распределения результатов наблюдений принадлежат тем или иным параметрическим семействам распределений, стала заметной частью статистики лишь со второй трети ХХ века. В 30-е годы появились работы А.Н.Колмогорова и Н.В.Смирнова, предложивших и изучивших статистические критерии, носящие в настоящее время их имена (история этих работ подробно описана в статье [30]). Эти критерии основаны на использовании так называемого эмпирического процесса - разности между эмпирической и теоретической функциями распределения, умноженной на квадратный корень из объема выборки. В работе А.Н.Колмогорова 1933 г. изучено предельное распределение супремума модуля эмпирического процесса, называемого сейчас критерием Колмогорова. Затем Н.В.Смирнов исследовал супремум и инфимум эмпирического процесса, а также интеграл (по теоретической функции распределения) квадрата эмпирического процесса.

Следует отметить, что встречающееся иногда в литературе [31] словосочетание "критерий Колмогорова-Смирнова", как подробно обосновано в статье [30], некорректно, поскольку эти два статистика никогда не печатались вместе и не изучали один и тот же критерий. Корректно сочетание "критерий типа Колмогорова-Смирнова", применяемое для обозначения критериев, основанных на использовании супремума функций от эмпирического процесса

После второй мировой войны развитие непараметрической статистики пошло быстрыми темпами. Большую роль сыграли работы Вилкоксона и его школы. К настоящему времени с помощью непараметрических методов можно решать практически тот же круг статистических задач, что и с помощью параметрических [31]. Все большую роль играют непараметрические оценки плотности, непараметрические методы регрессии и распознавания образов (дискриминантного анализа). В нашей стране непараметрические методы получили достаточно большую известность после выхода в 1965 г. первого издания упомянутого выше сборника статистических таблиц Л.Н.Большева и Н.В.Смирнова [25], содержащего подробные таблицы для основных непараметрических критериев.

Тем не менее параметрические методы всё еще популярнее непараметрических, особенно среди тех прикладников, кто слабо знаком со статистическими методами. Неоднократно публиковались (см. сводки в [4, 32]) экспериментальные данные, свидетельствующие о том, что распределения реально наблюдаемых случайных величин, в частности, ошибок измерения, в подавляющем большинстве случаев отличны от нормальных (гауссовских). Тем не менее теоретики продолжают строить и изучать статистические модели, основанные на гауссовости, а практики - применять подобные методы и модели. Другими словами, "ищут под фонарем, а не там, где потеряли".

**6. Устойчивость статистических процедур (робастность)**

Если в параметрических постановках на данных накладываются слишком жесткие требования - их функции распределения должны принадлежать определенному параметрическому семейству, то в непараметрических, наоборот, излишне слабые - требуется лишь, чтобы функции распределения были непрерывны. При этом игнорируется априорная информация о том, каков "примерный вид" распределения. Априори можно ожидать, что учет этого "примерного вида" улучшит показатели качества статистических процедур. Развитием этой идеи является теория устойчивости (робастности) статистических процедур, в которой предполагается, что распределение исходных данных мало отличается от некоторого параметрического семейства. С 60-х годов эту теорию разрабатывали П.Хубер[33], Ф.Хампель [34] и многие другие. Из монографий на русском языке, трактующих о робастности и устойчивости статистических процедур, самой ранней и наиболее общей была книга [35], следующей - монография [36]. Частными случаями реализации идеи робастности (устойчивости) статистических процедур являются рассматриваемые ниже статистика объектов нечисловой природы и интервальная статистика.

Имеется большое разнообразие моделей робастности в зависимости от того, какие именно отклонения от заданного параметрического семейства допускаются. Наиболее популярной [33,34] оказалась модель выбросов, в которой исходная выборка "засоряется" малым числом "выбросов", имеющих принципиально иное распределение. Однако эта модель представляется "тупиковой", поскольку в большинстве случаев большие выбросы либо невозможны из-за ограниченности шкалы прибора, либо от них можно избавиться, применяя лишь статистики, построенные по центральной части вариационного ряда. Кроме того, в подобных моделях обычно считается известной частота засорения, что в сочетании со сказанным выше делает их малопригодными для практического использования.

Более перспективным представляется модель Ю.Н.Благовещенского [37], в которой расстояние между распределением каждого элемента выборки и базовым распределением не превосходит заданной малой величины.

**7. Бутстреп (размножение выборок)**

Другое из упомянутых выше направлений - бутстреп - связано с интенсивным использованием возможностей вычислительной техники. Основная идея состоит в том, чтобы теоретическое исследование заменить вычислительным экспериментом. Вместо описания выборки распределением из параметрического семейства строим большое число "похожих" выборок, т.е. "размножаем" выборку. Затем вместо оценивания характеристик и параметров и проверки гипотез на основе свойств теоретического распределения решаем эти задачи вычислительным методом, рассчитывая интересующие нас статистики по каждой из "похожих" выборок и анализируя полученные при этом распределения. Например, вместо того, чтобы теоретическим путем находить распределение статистики, доверительные интервалы и другие характеристики, моделируют много выборок, похожих на исходную, рассчитывают соответствующие значения интересующей исследователя статистики и изучают их эмпирическое распределение. Квантили этого распределения задают доверительные интервалы, и т.д.

Термин "бутстреп" мгновенно получил известность после первой же статьи Б.Эфрона 1979 г. [39] по этой тематике. Он сразу же стал обсуждаться в массе публикаций, в том числе и научно-популярных [40]. В "Заводской лаборатории" была помещена подборка статей по бутстрепу [41], выпущен сборник статей Б.Эфрона [42]. Основная идея бутстрепа по Б.Эфрону состоит в том, что методом Монте-Карло (статистических испытаний) многократно извлекаются выборки из эмпирического распределения. Эти выборки, естественно, являются вариантами исходной, напоминают ее.

Сама по себе идея "размножения выборок" была известна гораздо раньше. Статья Б.Эфрона [39] называется так: "Бутстреп-методы: новый взгляд на метод складного ножа". Упомянутый "метод складного ножа" (jackknife) предложен М.Кенуем еще в 1949 г., за 30 лет до статьи Б.Эфрона. "Размножение выборок" при этом осуществляется путем исключения одного наблюдения. При этом для выборки объема n получаем n "похожих" на нее выборок объема (n - 1) каждая. Если же исключать по 2 наблюдения, то число "похожих" выборок возрастает до n (n - 1) / 2 объема (n - 2) каждая.

Преимущества и недостатки бутстрепа как статистического метода обсуждаются в [43]. Там же и в [18] приводится информация о ряде аналогичных методов. Необходимо подчеркнуть, что бутстреп по Эфрону [39-42] - лишь один из вариантов методов "размножения выборки" (resampling), и, на наш взгляд, не самый удачный. Метод "складного ножа" представляется более полезным. На его основе можно сформулировать следующую простую практическую рекомендацию.

Предположим, что Вы по выборке делаете какие-либо статистические выводы. Вы хотите узнать также, насколько эти выводы устойчивы. Если у Вас есть другие (контрольные) выборки, описывающие то же явление, то Вы можете применить к ним ту же статистическую процедуру и сравнить результаты. А если таких выборок нет? Тогда Вы можете их построить искусственно. Берете исходную выборку и исключаете один элемент. Получаете похожую выборку. Затем возвращаете этот элемент и исключаете другой. Получаете вторую похожую выборку. Поступив так со всеми элементами исходной выборки, получаете столько выборок, похожих на исходную, каков ее объем. Остается обработать их тем же способом, что и исходную, и изучить устойчивость получаемых выводов - разброс оценок параметров, частоты принятия или отклонения гипотез и т.д.

Можно изменять не выборку, а сами данные. Поскольку всегда имеются погрешности измерения, то реальные данные - это не числа, а интервалы (результат измерения плюс-минус погрешность). Нужна статистическая теория анализа таких данных.

**8. Статистика интервальных данных**

Перспективное и быстро развивающееся направление последних лет - математическая статистика интервальных данных. Речь идет о развитии методов математической статистики в ситуации, когда статистические данные - не числа, а интервалы, в частности, порожденные наложением ошибок измерения на значения случайных величин. Полученные результаты отражены, в частности, в выступлениях на проведенной в "Заводской лаборатории" дискуссии [44] и в докладах международной конференции ИНТЕРВАЛ-92 [45].

Статистика интервальных данных идейно связана с интервальной математикой, в которой в роли чисел выступают интервалы (см., например, монографию [46]). Это направление математики является дальнейшим развитием всем известных правил приближенных вычислений, посвященных выражению погрешностей суммы, разности, произведения, частного через погрешности тех чисел, над которыми осуществляются перечисленные операции. Как видно из докладов [45], к настоящему времени удалось решить, в частности, ряд задач теории интервальных дифференциальных уравнений, в которых коэффициенты, начальные условия и решения описываются с помощью интервалов.

Ведущая научная школа в области статистики интервальных данных - это школа проф.А.П.Вощинина, активно работающая с конца 70-х годов. Полученные результаты отражены в ряде монографий (см., в частности, [47,48,49]), статей [44], докладов [45], диссертаций [50,51]. В частности, изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности.

Рассмотрим другое направление в статистике интервальных данных, которое также представляется перспективным. В нем развиваются асимптотические методы статистического анализа интервальных данных при больших объемах выборок и малых погрешностях измерений. В отличие от классической математической статистики, сначала устремляется к бесконечности объем выборки и только потом - уменьшаются до нуля погрешности. В частности, с помощью такой асимптотики были сформулированы правила выбора метода оценивания параметров гамма-распределения в ГОСТ 11.011-83 [24].

В развитие идей, сформулированных в [52,53], разработана общая схема исследования, включающая расчет нотны (максимально возможного отклонения статистики, вызванного интервальностью исходных данных) и рационального объема выборки (превышение которого не дает существенного повышения точности оценивания). Она применена к оцениванию математического ожидания, дисперсии, коэффициента вариации, параметров гамма-распределения и характеристик аддитивных статистик, при проверке гипотез о параметрах нормального распределения, в т.ч. с помощью критерия Стьюдента, а также гипотезы однородности с помощью критерия Смирнова. Разработаны подходы к рассмотрению интервальных данных в основных постановках регрессионного, дискриминантного и кластерного анализов. В частности, изучено влияние погрешностей измерений и наблюдений на свойства алгоритмов регрессионного анализа, разработаны способы расчета нотн и рациональных объемов выборок, введены и исследованы новые понятия многомерных и асимптотических нотн, доказаны соответствующие предельные теоремы. Начата разработка интервального дискриминантного анализа, в частности, рассмотрено влияние интервальности данных на введенный нами показатель качества классификации. Изучено асимптотическое поведение оценок метода моментов и оценок максимального правдоподобия (а также более общих - оценок минимального контраста), проведено асимптотическое сравнение этих методов в случае интервальных данных. Найдены общие условия, при которых, в отличие от классической математической статистики, метод моментов дает более точные оценки, чем метод максимального правдоподобия (см. статью [54], в которой приведены также ссылки на другие публикации, относящиеся к рассматриваемому направлению в статистике интервальных данных)..

Как показала, в частности, международная конференция ИНТЕРВАЛ-92, в области асимптотической математической статистики интервальных данных российская наука имеет мировой приоритет. Развертывание работ по рассматриваемой тематике позволит закрепить этот приоритет, получить теоретические результаты, основополагающие в новой области математической статистики и необходимые для обоснованного статистического анализа почти всех типов данных. Со временем во все виды статистического программного обеспечения должны быть включены алгоритмы интервальной статистики, "параллельные" обычно используемым алгоритмам прикладной математической статистики. Это позволит в явном виде учесть наличие погрешностей у результатов наблюдений, сблизить позиции метрологов и статистиков.

**9. Статистика объектов нечисловой природы**

**как часть прикладной статистики**

Согласно классификации статистических методов, принятой в [55-56], прикладная статистика делится на следующие четыре области:

статистика (числовых) случайных величин,

многомерный статистический анализ,

статистика временных рядов и случайных процессов,

статистика объектов нечисловой природы.

Первые три из этих областей являются классическими. Остановимся на четвертой, только еще входящей в массовое сознание специалистов. Ее именуют также статистикой нечисловых данных или попросту нечисловой статистикой.

Исходный объект в математической статистике - это выборка. В вероятностной теории статистики выборка - это совокупность независимых одинаково распределенных случайных элементов. Какова природа этих элементов? В классической математической статистике (той, что обычно преподают студентам) элементы выборки - это числа. В многомерном статистическом анализе - вектора. А в нечисловой статистике элементы выборки - это объекты нечисловой природы, которые нельзя складывать и умножать на числа. Другими словами, объекты нечисловой природы лежат в пространствах, не имеющих векторной структуры.

Примерами объектов нечисловой природы являются (см. также обзоры [56-57]):

значения качественных признаков, т.е. результаты кодировки объектов с помощью заданного перечня категорий (градаций);

упорядочения (ранжировки) экспертами образцов продукции (при оценке ее технического уровня и конкурентоспособности)) или заявок на проведение научных работ (при проведении конкурсов на выделение грантов);

классификации, т.е. разбиения объектов на группы сходных между собой (кластеры);

толерантности, т.е. бинарные отношения, описывающие сходство объектов между собой, например, сходства тематики научных работ, оцениваемого экспертами с целью рационального формирования экспертных советов внутри определенной области науки;

результаты парных сравнений или контроля качества продукции по альтернативному признаку ("годен" - "брак"), т.е. последовательности из 0 и 1;

множества (обычные или нечеткие), например, зоны, пораженные коррозией, или перечни возможных причин аварии, составленные экспертами независимо друг от друга;

слова, предложения, тексты;

вектора, координаты которых - совокупность значений разнотипных признаков, например, результат составления статистического отчета о научно-технической деятельности (форма No.1-наука) или заполненная компьютеризированная история болезни, в которой часть признаков носит качественный характер, а часть - количественный;

ответы на вопросы экспертной, маркетинговой или социологической анкеты, часть из которых носит количественный характер (возможно, интервальный), часть сводится к выбору одной из нескольких подсказок, а часть представляет собой тексты; и т.д.

Интервальные данные (см. выше) тоже можно рассматривать как пример объектов нечисловой природы, а именно, как частный случай нечетких множеств.

С начала 70-х годов под влиянием запросов прикладных исследований в технических, медицинских и социально-экономических науках в России активно развивается статистика объектов нечисловой природы, известная также как статистика нечисловых данных или нечисловая статистика. В создании этой сравнительно новой области прикладной математической статистики приоритет принадлежит российским ученым.

Большую роль сыграл основанный в 1973 г. научный семинар "Экспертные оценки и анализ данных". В 60-е годы советское научное сообщество стало интересоваться методами экспертных оценок (об их истории и современном состоянии см. обзор [58]). Как следствие, началось знакомство с конкретными математизированными теориями, связанными с этими методами. Речь идет о репрезентативной теории измерений, ставшей известной в нашей стране по статье П.Суппеса и Дж.Зинеса в сборнике [59] и книге И.Пфанцагля [60], о теории нечеткости Л.А.Заде [61], теории парных сравнений, описанной в монографии Г.Дэвида [62]. К этому кругу идей примыкают теория случайных множеств (см., например, книгу Ж.Матерона [63]) и методы многомерного шкалирования (описаны в монографиях А.Ю.Терехиной [64] и В.Т.Перекреста [65]). Но наибольшее влияние оказали идеи Дж.Кемени, который аксиоматически ввел расстояние между ранжировками (теперь оно именуется в литературе расстоянием Кемени) и предложил использовать в качестве средней величины решение оптимизационной задачи (теперь - медиана Кемени). Его скромная книжка [66], написанная в соавторстве с Дж.Снеллом, породила большой поток исследований.

В течение 70-х годов на основе запросов теории экспертных оценок (а также социологии, экономики, техники и медицины) развивались конкретные направления статистики объектов нечисловой природы. Были установлены связи между конкретными видами таких объектов, разработаны для них вероятностные модели (см. обзор [67]). Итоги этого периода подведены в монографиях [35,68,69]).

Следующий этап - выделение статистики объектов нечисловой природы в качестве самостоятельного направления в прикладной статистике, ядром которого являются методы статистического анализа данных произвольной природы. Программа развития этого нового научного направления впервые была сформулирована в статье [70]. Реализация этой программы была осуществлена в 80-е годы. Для работ этого периода характерна сосредоточенность на внутренних проблемах нечисловой статистики. Ссылки на конкретные монографии, сборники, статьи и иные публикации нескольких десятков авторов приведены в обзорах [56,57,67]. Отметим лишь сборник [71], специально посвященный нечисловой статистике, и диссертацию [72], относящуюся к непараметрической теории парных сравнений.

К 90-м годам статистика объектов нечисловой природы с теоретической точки зрения была достаточно хорошо развита, основные идеи, подходы и методы были разработаны и изучены математически, в частности, доказано достаточно много теорем. Однако она оставалась недостаточно апробированной на практике. Это было связано как с ее сравнительной молодостью, так и с общеизвестными особенностями организации науки в 80-е годы, когда отсутствовали достаточные стимулы к тому, чтобы теоретики занялись широким внедрением своих результатов. И в 90-е годы наступило время от математико-статистических исследований перейти к применению полученных результатов на практике. Эта тенденция хорошо отражена в монографиях [73,74], материалах международной конференции "Управление большими системами" [75].

Следует отметить, что в статистике объектов нечисловой природы, как и в других областях прикладной математической статистики и прикладной математики вообще, одна и та же математическая схема может с успехом применяться и в технических исследованиях, и в медицине, и в социологии, и для анализа экспертных оценок, а потому ее лучше всего формулировать и изучать в наиболее общем виде, для объектов произвольной природы.

**10. Основные идеи статистики объектов нечисловой природы**

В чем принципиальная новизна нечисловой статистики? Для классической математической статистики характерна операция сложения. При расчете выборочных характеристик распределения (выборочное среднее арифметическое, выборочная дисперсия и др.), в регрессионном анализе и других областях этой научной дисциплины постоянно используются суммы. Математический аппарат - законы больших чисел, Центральная предельная теорема и другие теоремы - нацелены на изучение сумм. В нечисловой же статистике нельзя использовать операцию сложения, поскольку элементы выборки лежат в пространствах, где нет операции сложения. Методы обработки нечисловых данных основаны на принципиально ином математическом аппарате - на применении различных расстояний в пространствах объектов нечисловой природы.

Кратко рассмотрим несколько идей, развиваемых в статистике объектов нечисловой природы для данных, лежащих в пространствах произвольного вида. Решаются классические задачи описания данных, оценивания, проверки гипотез - но для неклассических данных, а потому неклассическими методами.

Первой обсудим проблему определения средних величин. В рамках репрезентативной теории измерений удается указать вид средних величин, соответствующих тем или иным шкалам измерения [35]. В классической математической статистике средние величины вводят с помощью операций сложения (выборочное среднее арифметическое, математическое ожидание) или упорядочения (выборочная и теоретическая медианы). В пространствах произвольной природы средние значения нельзя определить с помощью операций сложения или упорядочения. Теоретические и эмпирические средние приходится вводить как решения экстремальных задач. Для теоретического среднего это - задача минимизации математического ожидания (в классическом смысле) расстояния от случайного элемента со значениями в рассматриваемом пространстве до фиксированной точки этого пространства (минимизируется указанная функция от этой точки). Для эмпирического среднего математическое ожидание берется по эмпирическому распределению, т.е. берется сумма расстояний от некоторой точки до элементов выборки и затем минимизируется по этой точке. При этом как эмпирическое, так и теоретическое средние как решения экстремальных задач могут быть не единственным элементом пространства, а состоять из множества таких элементов, которое может оказаться и пустым. Тем не менее удалось сформулировать и доказать законы больших чисел для средних величин, определенных указанным образом, т.е. установить сходимость эмпирических средних к теоретическим .

Оказалось, что методы доказательства законов больших чисел допускают существенно более широкую область применения, чем та, для которой они были разработаны. А именно, удалось изучить асимптотику решений экстремальных статистических задач, к которым, как известно, сводится большинство постановок прикладной статистики [26]. В частности, кроме законов больших чисел установлена и состоятельность оценок минимального контраста, в том числе оценок максимального правдоподобия и робастных оценок. К настоящему времени подобные оценки изучены также и в интервальной статистике.

В статистике в пространствах произвольной природы большую роль играют непараметрические оценки плотности, используемые, в частности, в различных алгоритмах регрессионного, дискриминантного, кластерного анализов. В нечисловой статистике предложен и изучен ряд типов непараметрических оценок плотности в пространствах произвольной природы, в частности, доказана их состоятельность, изучена скорость сходимости и установлен примечательный факт совпадения наилучшей скорости сходимости в произвольном случае с той, которая имеет быть в классической теории для числовых случайных величин.

Дискриминантный, кластерный, регрессионный анализы в пространствах произвольной природы основаны либо на параметрической теории - и тогда применяется подход, связанный с асимптотикой решения экстремальных статистических задач - либо на непараметрической теории - и тогда используются алгоритмы на основе непараметрических оценок плотности.

Для проверки гипотез могут быть использованы статистики интегрального типа, в частности, типа омега-квадрат. Любопытно, что предельная теория таких статистик, построенная первоначально в классической постановке [76], приобрела естественный (завершенный, изящный) вид именно для пространств произвольного вида [77], поскольку при этом удалось провести рассуждения, опираясь на базовые математические соотношения, а не на те частные (с общей точки зрения), что были связаны с конечномерным пространством.

Представляют интерес результаты, связанные с конкретными областями статистики объектов нечисловой природы, в частности, со статистикой нечетких множеств [78], со случайными множествами [35] (следует отметить, что теория нечетких множеств в определенном смысле сводится к теории случайных множеств [35,78]), с непараметрической теорией парных сравнений [72], с аксиоматическим введением метрик в конкретных пространствах объектов нечисловой природы [71].

Для анализа нечисловых, в частности, экспертных данных весьма важны методы классификации. С другой стороны, наиболее естественно ставить и решать задачи классификации, основанные на использовании расстояний или показателей различия, в рамках статистики объектов нечисловой природы. Это касается как распознавания образов с учителем (другими словами, дискриминантного анализа), так и распознавания образов без учителя (т.е. кластерного анализа). Современное состояние дискриминантного и кластерного анализа с точки зрения статистики объектов нечисловой природы отражено работах в [79].

Статистические методы анализа нечисловых данных особенно хорошо приспособлены для применения в экономике, социологии и экспертных оценках, поскольку в этих областях от 50% до 90% данных являются нечисловыми [80].

**11. Заключение**

Мы рассмотрели пять "точек роста" прикладной статистики как методической дисциплины. Разумеется, они не исчерпывают все многообразие фронта научных исследований в этой области. В частности, решены отнюдь не все проблемы, поставленные в конце 70-х годов в т.н. "цахкадзорской тетради" [81]. Кроме того, мы почти не затрагивали разнообразные применения статистических методов в конкретных прикладных областях. Много интересных проблем есть в планировании экспериментов, особенно кинетических (см., например, статьи [2,82]), при анализе проблем надежности (см., в частности, статью [83]), в новых статистических методах управления качеством продукции, в том числе в связи с идеями Г.Тагути (см. об этом статью [84]), в вопросах экологии и безопасности [75], и др.

В течение последних более чем 60 лет в России наблюдается огромный разрыв между государственной статистикой и научным сообществом специалистов по статистическим методам (подробнее об этом см.[19]). В учебнике по истории статистики [14] даже не упоминаются имена членов-корреспондентов АН СССР Н.В.Смирнова и Л.Н.Большева! Поэтому нет ничего удивительного в том, что тенденции развития современной прикладной математической статистики столь же мало обсуждаются отечественными авторами, как и ее история. Буду рад, если настоящая статья положит начало дискуссии о будущем нашей науки.

**ЛИТЕРАТУРА**

1. Орлов А.И. / Вестник статистики. 1986, № 8. С.52 - 56

2. Горский В.Г. - В сб.: Международная школа повышения квалификации "Инженерно-химическая наука для передовых технологий". Труды третьей сессии, 26-30 мая 1997. Казань, Россия / Под ред. В.А.Махлина. - М.: Научно-Исследовательский Физико-Химический Институт им.Карпова, 1997. С.261-293.

3. Гуда А.Н. Модели, методы и средства анализа данных в затрудненных условиях. Автореф. дисс. докт. технич. наук. - Таганрог: Таганрогский государственный радиотехнический университет, 1997. 38 с.

4. Налимов В.В. Применение математической статистики при анализе вещества. - М.: Физматгиз, 1960. - 430 с.

5. Налимов В.В., Чернова Н.Л. Статистические методы планирования экстремальных экспериментов. - М.: Физматгиз, 1965. - 340 с.

6. Налимов В.В. Канатоходец. Воспоминания. - М.: Издательская группа "Прогресс", 1994. - 456 с.

7. Гнеденко Б.В., Орлов А.И. / Заводская лаборатория. 1988. Т.54. № 1. С.1-4.

8. Горский В.Г. / Заводская лаборатория. 1992. Т.58. № 1. С.63-64.

9. Орлов А.И. / Заводская лаборатория. 1992. Т.58. № 1. С.67-74.

10. Комаров Д.М., Орлов А.И. - В сб.: Вопросы применения экспертных систем. - Минск: Центросистем, 1988. С.151-160.

11. Ленин В.И. Развитие капитализма в России. Процесс образования внутреннего рынка для крупной промышленности. - М.: Политиздат, 1986. - XII, 610 с.

12. Гнеденко Б.В. Курс теории вероятностей: Учебник. - Изд. 6-е,

перераб. и доп. - М.: Наука, Гл. ред. физ.-мат. лит., 1988. - 448 с.

13. Клейн Ф. Лекции о развитии математики в 19 столетии. Часть I. - М.

-Л.: Объединенное научно-техническое издательство НКТП СССР, 1937. - 432 с.

14. Плошко Б.Г., Елисеева И.И. История статистики: Учеб. пособие. - М.:

Финансы и статистика. 1990. - 295 с.

15. Гнеденко Б.В. Математическая статистика и контроль качества. - М.: Знание, 1976. - 64 с.

16. Орлов А.И. / Заводская лаборатория. 1997. Т.63. № 3. С.55-62.

17. Бернштейн С.Н. В сб.: Труды Всероссийского съезда математиков в Москве 27 апреля - 4 мая 1927 г. - М.-Л.: ГИЗ, 1928. С.50-63.

18. Орлов А.И. / Надежность и контроль качества. 1987. № 6. С. 54-59.

19. Орлов А.И. / Вестник статистики. 1990, № 1. С.65 - 71.

20. Кендалл М., Стьюарт А. Теория распределений. - М.: Наука, 1966. 566 с.

20. Кендалл М., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. 899 с.

20. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. 736 с.

23. Налимов В.В., Мульченко З.М. Наукометрия. Изучение развития науки как информационного процесса. - М.:Наука, 1969. 192 с.

24. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения. - М.: Изд-во стандартов. 1984. - 53 с.

25. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1965 (1-е изд.), 1968 (2-е изд.), 1983 (3-е изд.).

26. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. - М.: Финансы и статистика, 1983. - 471 с.

27. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. - М.: Финансы и статистика, 1985. - 488 с.

28. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. - М.: Финансы и статистика, 1989. - 607 с.

29. Орлов А.И. / Надежность и контроль качества. 1991. № 8. С.3-8.

30. Орлов А.И. / Заводская лаборатория. 1995. Т.61. № 7. С.59-61.

31. Холландер М., Вулф Д. Непараметрические методы статистики. - М.: Финансы и статистика, 1985.

32. Орлов А.И. /Заводская лаборатория. 1991. Т.57. № 7. С.64-66.

33. Хьюбер П. Робастность в статистике. - М.: Мир, 1984. - 304 с.

34. Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния. - М.: Мир, 1989. - 512 с.

35. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.

36. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: Статистическая обработка неоднородных совокупностей. - М;: Статистика, 1980. - 208 с.

37. Благовещенский Ю.Н. - В сб.: Тезисы докладов Международной конференции по теории вероятностей и математической статистике. Вильнюс, 25-30 июня 1973 г. Т.1. - Вильнюс: Изд-во Вильнюсского госуниверситета, 1973. С.77-78.

39. Efron B. / Ann. Statist. 1979. V.7. № 1. P.1-26.

40. Диаконис П., Эфрон Б. / В мире науки. 1983. № 7. С.60-73.

41. Подборка статей по бутстрепу / Заводская лаборатория. 1987. Т.53. № 10. С.76-99.

42. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. - М.: Финансы и статистика, 1988. 263 с.

43. Орлов А.И. /Заводская лаборатория. 1987. Т.53. № 10. С.82-85.

44. Дискуссия по анализу интервальных данных / Заводская лаборатория. 1990. Т.56. № 7. С.75-95.

45. Сборник трудов Международной конференции по интервальным и стохастическим методам в науке и технике. Тт. 1,2. - М.: МЭИ, 1992.

46. Шокин Ю.И. Интервальный анализ. - Новосибирск: Наука, 1981. 284 с.

47. Вощинин А.П. Метод оптимизации объектов по интервальным моделям целевой функции. - М.: МЭИ, 1987. 109 с.

48. Вощинин А.П., Сотиров Г.Р. Оптимизация в условиях неопределенности. - М.: МЭИ - София: Техника, 1989. 224 с.

49. Вощинин А.П., Акматбеков Р.А. Оптимизация по регрессионным моделям и планирование эксперимента. - Бишкек: Илим, 1991. 164 с.

50. Дывак Н.П. Разработка методов оптимального планирования эксперимента и анализа интервальных данных. Автореф. дисс. канд.. технич. наук. - М.: МЭИ, 1992. 20 с.

51. Симов С.Ж. Разработка и исследование интервальных моделей при анализе данных и проектировании экспертных систем. Автореф. дисс. канд.. технич. наук. - М.: МЭИ, 1992. 20 с.

52. Орлов А.И. / Заводская лаборатория. 1990, Т.56. № 7. С.86-89.

53. Orlov A.I. / Interval Computations. 1992. № 1(3). Р.44-52.

54. Орлов А.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1995. С. 114-124.

55. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. - М.: ВНИИСтандартизации, 1987. - 64 с.

56. Орлов А.И. / Заводская лаборатория. 1990. Т.56. № 3. С.76-83.

57. Орлов А.И. / Заводская лаборатория. 1995. Т.61. № 3. С.43-52.

58. Орлов А.И. / Заводская лаборатория. 1996. Т.62. № 1. С.54-60.

59. Суппес П., Зинес Дж. - В сб.: Психологические измерения. -М: Мир,1967. С. 9-110.

60. Пфанцагль И. Теория измерений. - М.: Мир, 1976. 166 с.

61. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. - М.: Мир, 1976. 168 с.

62. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978. 144 с.

63. Матерон Ж. Случайные множества и интегральная геометрия. - М.: Мир, 1978. 318 с.

64. Терехина А.Ю. Анализ данных методами многомерного шкалирования. - М.: Наука, 1986. 168 с.

65. Перекрест В.Т. Нелинейный типологический анализ социально-экономической информации: Математические и вычислительные методы. - Л.: Наука, 1983. 176 с.

66. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. - М.: Советское радио, 1972. 192 с.

67. Орлов А.И. / Заводская лаборатория. 1995. Т.61. № 5. С.43-51.

68. Тюрин Ю.Н., Литвак Б.Г., Орлов А.И., Сатаров Г.А., Шмерлинг Д.С. Анализ нечисловой информации. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1981. - 80 с.

69. Литвак Б.Г. Экспертная информация: Методы получения и анализа. - М.: Радио и связь, 1982. 184 с.

70. Орлов А.И. - В сб.: Экспертные оценки. Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. С.17-33.

71. Анализ нечисловой информации в социологических исследованиях. - М.: Наука, 1985. - 220 с.

72. Рыданова Г.В. Некоторые вопросы статистического анализа случайных бинарных векторов. Автореф. дисс. канд. физ.-мат. наук. - М.: МГУ, 1988. 16 с.

73. Ларичев О.И., Мошкович Е.М. Качественные методы принятия решений. Вербальный анализ решений. - М.: Наука, 1996. 208 с.

74. Литвак Б.Г. Экспертные оценки и принятие решений. - М.: Патент, 1996. 271 с.

75. Управление большими системами. Материалы международной научно-практической конференции (22-26 сентября 1997 г., Москва, Россия). Общая редакция - Бурков В.Н., Новиков Д.А. - М.: СИНТЕГ, 1997. 432 с.

76. Орлов А.И. / Доклады АН СССР. 1974. Т.219. № 4. С.808-811.

77. Орлов А.И. - В сб.: Вероятностные процессы и их приложения. - М.: МИЭМ, 1989. С.118-123.

78. Орлов А.И. Задачи оптимизации и нечеткие переменные. - М.: Знание, 1980. - 64 с.

79. Орлов А.И. / Социология: методология, методы, математические модели. 1992. № 2. С.28-50.

80. Титма М.Х., Тоодинг Л.М. / Социологические исследования. 1986. № 4. С.123-128.

81. Загоруйко Н.Г., Орлов А.И. - В сб.: Современные проблемы кибернетики (прикладная статистика). - М.: Знание, 1981. С.53-63.

82. Горский В.Г., Денисов В.И., Иткина Н.Б. / Заводская лаборатория. 1992. Т.58. № 1. С.64-67.

83. Тескин О.И. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1995. С. 227 - 236.

84. Адлер Ю.П., Талалай А.М. / Курс на качество. 1992. № 3-4. С.85-93.