**ОБЪЕКТЫ НЕЧИСЛОВОЙ ПРИРОДЫ.**

Статистика объектов нечисловой природы - это направление в математической статистике, в котором в качестве статистических данных (результатов наблюдений) рассматриваются объекты нечисловой природы. Так называют объекты, которые нецелесообразно описывать числами, в частности элементы нелинейных пространств. Примерами являются бинарные отношения (ранжировки, разбиения, толерантности и др.), результаты парных и множественных сравнений, множества, нечеткие множества, измерение в шкалах, отличных от абсолютных. Этот перечень примеров не претендует на законченность. Он складывался постепенно в соответствии с исследованиями в области статистики объектов нечисловой природы.

Объекты нечисловой природы широко используются в теоретических и прикладных исследованиях по проблемам управления, в частности управления качеством продукции, в технических науках, медицине, социологии, экономике, психологии и т.д., а также практически во всех отраслях народного хозяйства. Это обосновывает целесообразность дальнейшего развития рассматриваемого математико-статистического аппарата.

В журнале " Заводская лаборатория " опубликовано два обзора [1,2] и несколько десятков статей ([3-33] и др.) по различным вопросам статистики объектов нечисловой природы. Однако литературы, в которой бы разъяснялись основные понятия этого направления анализа статистических данных, явно не хватает. Данный обзор и посвящен первоначальному знакомству с основными видами объектов нечисловой природы.

**Основные понятия репрезентативной теории измерений**

Начнем с рассмотрения конкретного социологического исследования. Обсуждение можно вести в терминах экспертных оценок. Тогда вместо сравнения математики и физики *n* экспертов (а не выпускников школ) оценивают по конкурентоспособности на мировом рынке, например, две марки стали. Однако в настоящее время социологические исследования более привычны, чем экспертные.

При изучении привлекательности различных профессий для выпускников новосибирских школ [34] был составлен список из 30 профессий. Опрашиваемых просили оценить каждую из этих профессий одним из баллов 1,2,...,10 по правилу: чем больше нравится, тем выше балл. Для получения социологических выводов необходимо было дать единую оценку привлекательности определенной профессии для совокупности выпускников школ. В качестве такой оценки в работе [34] использовалось среднее арифметическое баллов, выставленных профессии опрошенными школьниками. В частности, физика получила средний балл 7.69, а математика - 7.50. В соответствии с логикой [34], физика более предпочтительна, чем математика.

Однако было отмечено [35], что этот вывод противоречит данным работы [36], согласно которым ленинградские школьники средних классов больше любят математику, чем физику. Обсудим одно из возможных объяснений этого противоречия, которое заключается в неадекватной методике обработки данных, применных в работе [34].

Дело в том, что баллы 1,2,...,10 введены исследователем-социологом, т.е. субъективно. Если одна профессия оценена в 10 баллов, а вторая - в 2, то из этого нельзя заключить, что первая ровно в 5 раз привлекательней другой. Другой коллектив социологов мог бы принять иную систему баллов, например 1,4,9,16,...,100. Естественно предположить , что упорядочивание профессий по привлекательности, присущее школьникам, не зависит от того, какой системой баллов им предложит пользоваться социолог. Раз так, то распределение профессий по градациям десятибалльной системы не изменится, если перейти к другой системе баллов с помощью строго возрастающей функции .Если -ответы *n* выпускников школ, касающихся математики, а -физики, то после перехода к новой системе баллов ответы относительно математики будут иметь вид ,а относительно физики - .



Пусть единая оценка привлекательности профессии вычисляется с помощью функции. . Какие требования естественно наложить на функцию , чтобы полученные с ее помощью выводы не зависели от того, какой именно системой баллов пользовался социолог ?



Единая оценка вычислялась для того, чтобы сравнивать профессии по привлекательности. Поэтому потребуем устойчивости результата сравнения: неравенство



(1)

справедливо тогда и только тогда, когда справедливо неравенство

, (2)



причем равносильность неравенств (1) и (2) имеется при любых и . Какие устойчивы относительно сравнения? Ответ на этот вопрос был дан в работе [35]. В частности, оказалось, что средним арифметическим, как в работе [34], пользоваться нельзя, а членами вариационного ряда (и только ими) - можно.



Выше показан переход от социологической задачи к математической (подробнее см. [37, § 3.1], [38]), а именно, к одной из частных постановок проблемы адекватности в репрезентативной теории измерения [39, 40]. Изложим основные понятия этой теории в модификации , данной в работе [ 37 гл. 3 ].

Исходным понятием является совокупность Ф={}. допустимых преобразований шкалы (обычно Ф- группа),. Алгоритм обработки данных *а* , т.е. функция .(А - множество возможных результатов работы алгоритма) называется в шкале совокупностью допустимых преобразований Ф адекватным [ 37, c. 97], если



(3)



для всех , i=1,...,n и всех Ф. Таким образом, теорию измерений рассматриваем как теорию инвариантов относительно различных совокупностей допустимых преобразований Ф. Интерес вызывают две задачи: а) дано Ф; какие *а* из определенного класса удовлетворяют условию (3); б) дан алгоритм *а*; для каких Ф справедливо условие (3)? Уточнение этих постановок дано в работах [ 37, 41, 42 ]



Наиболее распространенные шкалы измерения описываются с помощью групп допустимых преобразований Ф. Если Ф состоит из всех взаимнооднозначных преобразований , то измерения проводятся в шкале наименований. Для порядковой шкалы Ф состоит из всех строго возрастающих преобразований. по этим двум шкалам измеряются качественные признаки.

Для шкалы интервалов Ф={ах+b; а>0,}, для шкалы отношений Ф=ах;а>0}, для шкалы разностей Ф={х+b;},для абсолютной шкалы Ф={, По этим четырем шкалам измеряются количественные признаки. В абсолютной шкале известно начало отсчета и единица измерения, в шкале отношений фиксированно начало , но не единица измерения, в шкале разностей, наоборот, единица измерения фиксирована, а начало отсчета - нет , в шкале интервалов ни то, ни другое не задано.



Различные свойства шкал, примеры реальных величин, измеряемых по тем или иным шкалам, приведены в работах [37, 39, 40,43].

**Бинарные отношения.**

Пусть *а*: - адекватный алгоритм в шкале наименований. Легко видеть [37,c.109] , что *a* - есть функция от матрицы **B**==**В**(), где



.



Если *a*: - адекватный алгоритм в шкале порядка, то *a* есть [37,c.111] функция от матрицы **C**==**C**() порядка n x n, где



Матрицы **B** и **C** можно проинтерпретировать в терминах бинарных отношений. Пусть некоторая характеристика измеряется у n объектов , причем - результат ее измерения у объекта Тогда матрицы B и C задают бинарные отношения на множестве объектов Q ={}Поскольку бинарное отношение можно рассматривать как подмножество декартова квадрата Q x Q, то любой матрице **D** = порядка n x n из 0 и 1 соответствует бинарное соотношение R(**D**), определяемое следующим образом: R(**D**) тогда и только тогда, когда 1.



Бинарное отношение R(**B**) - отношение эквивалентности, т.е. рефлексивное симметричное транзитивное отношение. Оно задает разбиение Q на классы эквивалентности. Два объекта и входят в один класс эквивалентности тогда и только тогда, когда .



Выше показано, как разбиения возникают в результате измерений в шкале наименований. Разбиения могут появляться и непосредственно. Так, при оценке качества промышленной продукции эксперты дают разбиение показателей качества на группы [44]. Для изучения психологического состояния людей их просят разбить предъявленные рисунки на группы сходных между собой [45,46].Аналогичная методика применяется в экспериментальных психологических исследованиях. [47,48].

Во многих задачах прикладной статистики разбиения получаются "на выходе" (в кластер-анализе) или же используются на промежуточных этапах анализа данных (например, сначала проводят классификацию с целью выделения однородных групп, а затем в каждой группе строят регрессионную зависимость, как в работе [49]).

Бинарное отношение R(**С**) задает разбиение Q на классы эквивалентности, между которыми введено отношение строгого порядка. Два объекта и входят в один класс тогда и только тогда, когда = 1 и = 1, т.е. Класс эквивалентности предшествует классу эквивалентности тогда и только тогда, когда для любых , имеем , = 1,= 0, т.е.. Такое бинарное отношение в статистике называют ранжировкой со связями [50]; связанными считаются объекты, входящие в один класс эквивалентности. В литературе встречаются и другие названия: линейный квазипорядок [51], упорядочение [52,гл.2], квазисерия [53, с.37]. Если каждый из классов эквивалентности состоит только из одного элемента, то имеем обычную ранжировку (другими словами, линейный порядок).



Как известно, ранжировки возникают в результате измерений в порядковой шкале. Так, при описанном выше опросе ответ выпускника школы - это ранжировка (со связями) профессий по привлекательности. Ранжировки часто возникают и непосредственно, без промежуточного этапа - приписывания объектам квазичисловых оценок - баллов. Многочисленные примеры тому даны М.Кендэлом [50]. При оценке качества промышленной продукции нормативные методические документы предусматривают использование ранжировок [44].

Для прикладных областей, кроме ранжировок и разбиений, представляют интерес толерантности, т.е. рефлексивные симметричные отношения [54]. Толерантность - математическая модель для выражения представлений о сходстве (похожести, близости). Разбиения - частный вид толерантностей. Однако в общем случае толерантность не обязана быть транзитивной. Необходимость использования толерантностей показана Э.Борелем при обсуждении физической непрерывности согласно Пуанкаре [55, с.88-91]. Толчок к более подробному изучению толерантностей дали исследования деятельности мозга [56]. Толерантности появляются и в других постановках, например, как результат парных сравнений (см.ниже).

Напомним, что любое бинарное отношение на конечном множестве может быть описано матрицей из 0 и 1.

**Дихотомические данные.**

Это данные , которые могут принимать одно из двух значений (0 или 1), т.е. результаты измерений альтернативного признака. Как уже было показано, измерения в шкале наименований и порядковой шкале приводят к бинарным отношениям, а те могут быть выражены как результаты измерений по нескольким альтернативным признакам, соответствующим элементам матриц, описывающих отношения. Дихотомические данные возникают в прикладных исследованиях и многими иными путями.

В настоящее время в большинстве стандартов на конкретную продукцию предусмотрен контроль по альтернативному признаку. Обширные теоретические исследования проблем статистического приемочного контроля по альтернативному признаку [57,58]. Основополагающими в этой области являются работы А.Н.Колмогорова [59,60]. Подход советской вероятностно-статистической школы к проблемам качества продукции по альтернативному признаку означает, что единица продукции относится к одной из двух категорий - "годных" или "дефектных", т.е. соответствующих или не соответствующих требованиям стандарта.

Дихотомические данные - давний объект математической статистики (см., например, [62, гл.33]) Особенно большое применение они имеют в медико-биологических [46] и социологических [63] исследованиях, в которых большинство переменных, интересующих специалистов, не может быть измерено ( в настоящее время!) по количественным шкалам. При этом дихотомические данные зачастую являются более адекватными, чем результаты измерений по методикам, использующим большее число градаций. В частности, психологические тесты типа MMPI [45] используют только дихотомические данные. На них опираются и методы парных сравнений [64].

Элементарным актом в методике парных сравнений является предъявление эксперту для сравнения двух объектов ( сравнение может проводиться также прибором). В одних постановках эксперт должен выбрать из двух объектов лучший по качеству, в других - ответить, похожи объекты или нет. В обоих случаях ответ эксперта можно выразить одной из двух цифр - 0 или 1. В первой постановке: 0, если лучшим объявлен первый объект; 1 - если второй. Во второй постановке: 0, если объекты похожи, схожи, близки; 1 - в противном случае.

Подводя итоги изложенному, можно сказать, что рассмотренные выше данные представимы в виде векторов из 0 и 1 ( при этом матрицы, очевидно, могут быть записаны в виде векторов). С.А.Айвазян [65] предлагает "унифицированную форму записи наблюдений", в которой любые виды результатов записываются в виде векторов из 0 и 1. Представляется, что это предложение имеет скорее академический интерес, но во всяком случае можно констатировать, что анализ дихотомических данных необходим во многих прикладных постановках.

**Множества**

Совокупность векторов X = () из 0 и 1 размерности n находится во взаимно-однозначном соответствии с совокупностью всех подмножеств множества N = {1, 2, ..., n}. При этом вектору X = () соответствует подмножество N(X)N, состоящее из тех и только из тех i, для которых = 1. Это объясняет, почему изложение вероятностных и статистических результатов, относящихся к анализу данных, являющихся объектами нечисловой природы перечисленных выше видов, велось [37, гл.4] на языке конечных случайных множеств.



Множества как исходные данные появляются и в иных постановках. Из геологических реалий исходил Ж.Матерон [66], из электротехнических - Н.Н.Ляшенко [67] и др. Случайные множества применялись для описания процесса случайного распространения, например распространения эпидемии или пожара[68, 69, 70] , а также в математической экономике [71]. Много работ связано с изучением случайных геометрических объектов - точек, прямых, кругов, мозаик и т.д. (обзор по состоянию на 1969г. дан а работе [72]). В работе [37, §4.6, §5.6] рассмотрены приложения случайных множеств в теории экспертных оценок и в теории управления запасами и ресурсами.

Отметим, что реальные объекты можно моделировать случайными множествами как из конечного числа элементов, так и из бесконечного, однако при расчетах на ЭВМ неизбежна дискретизация, т.е. переход к первой из названных возможностей.

**Нечеткие множества**

Пусть A - некоторое множество. Подмножество B множества A характеризуется своей характеристической функцией

(4)



Нечеткое подмножество множества характеризуется своей функцией принадлежности .. имеет вид (4) при некотором , то есть обычное (четкое) подмножество A.



Обычное подмножество можно было бы отождествить с его характеристической функцией. Этого не делают, поскольку для задания функции (в ныне принятом подходе) необходимо сначала задать множество. Нечеткое же подмножество с формальной точки зрения можно отождествить с его функцией принадлежности. Однако термин "нечеткое подмножество" предпочтительнее при построении математических моделей реальных явлений.

Начало современной теории нечеткости положено статьей Л.А.Заде в 1965г [73]. К настоящему времени по этой теории опубликованы тысячи книг и статей, издается несколько международных журналов, выполнено достаточно много как теоретических, так и прикладных работ. Из публикаций на русском языке, кроме перевода монографии Л.А.Заде, назовем книги С.А.Орловского [75], В.Б.Кузьмина [76], а также работы [77-80].

Л.А.Заде рассматривал теорию нечетких множеств как аппарат анализа и моделирования гуманистических систем, т.е. систем, в которых участвует человек [81, с.6]. Его подход "опирается на предпосылку о том, что элементами мышления человека являются не числа, а элементы некоторых нечетких множеств или классов объектов, для которых переход от "принадлежности" к "непринадлежности" не скачкообразен, а непрерывен " [81, с.7]. В настоящее время методы теории нечеткости используются почти во всех прикладных областях, в том числе при управлении качеством продукции и технологическими процессами. Популярный обзор прикладных возможностей теории нечеткостей дан в работах [43, 82].

Пусть и - два нечетких подмножества с функциями принадлежности и соответственно. Пересечением , произведением , объединением , отрицанием , суммой называются нечеткие подмножества с функциями принадлежности



соответственно.

Свойства введенных операций над нечеткими множествами и их связь с операциями над обычными множествами обсуждаются в работах [37,43].

**Объекты нечисловой природы как статистические данные**

В математической статистике наиболее распространенный объект изучения - выборка т.е. совокупность результатов наблюдений. В различных областях статистики результат наблюдения - это или число, или конечномерный вектор, или функция... Соответственно проводится деление математической статистики: одномерная статистика, многомерный статистический анализ, статистика временных рядов и случайных процессов... В статистике объектов нечисловой природы в качестве результатов наблюдений рассматриваются объекты нечисловой природы, в частности, перечисленных выше видов - измерения в шкалах, отличных от абсолютной, бинарные отношения, вектора из 0 и 1, множества, нечеткие множества. Выборка может состоять изранжировок и толерантностей, или множеств, или нечетких множеств и т.д.



Отметим необходимость развития методов статистической обработка "разнотипных данных", обусловленную большой ролью в прикладных исследованиях "признаков смешанной природы" [83]. Речь идет о том, что результат наблюдения состояния объекта зачастую представляет собой вектор, у которого часть координат измерена по шкале наименований, часть - по порядковой шкале, часть - по шкале интервалов и т.д. Статистические методы ориентированы обычно либо на абсолютную шкалу, либо на шкалу наименований (анализ таблиц сопряженности), а потому зачастую непригодны для обработки разнотипных данных. Есть и более сложные модели разнотипных данных, например, когда некоторые координаты вектора наблюдений описываются нечеткими множествами [43].

Для обозначения подобных неклассических результатов наблюдений в 1979 г. [84] предложен собирательный термин - объекты нечисловой природы. Термин "нечисловой" означает, что структура [85] пространства, в котором лежат результаты наблюдений, не является структурой действительных чисел, векторов или функций, она вообще не является структурой линейного (векторного) пространства. При расчетах объекты числовой природы, разумеется, изображаются с помощью чисел.

С целью "стандартизации математических орудий" [85, с.253] целесообразно разрабатывать методы статистического анализа данных, пригодные одновременно для всех перечисленных выше видов результатов наблюдений. Кроме того, в процессе развития прикладных исследований выявляется необходимость использования новых видов объектов нечисловой природы, отличных от рассмотренных выше, например, в связи с развитием статистических методов обработки текстовой информации [86].Поэтому целесообразно ввести еще один вид объектов нечисловой природы - объекты произвольной природы, т.е. элементы множества, на которые не наложено никаких условий (кроме "условий регулярности", необходимых для справедливости доказываемых теорем). Другими словами, в этом случае предполагается, что результаты наблюдений (элементы выборки) лежат в произвольном пространстве . Для получения теорем необходимо потребовать, чтобы удовлетворяло некоторым условиям, например, было топологическим пространством. Как известно, ряд результатов математической статистики получен именно в такой постановке. Так, при изучении оценок максимального правдоподобия элементы выборки могут лежать в пространстве произвольной природы. Это не влияет на рассуждения, поскольку в них рассматривается лишь зависимость плотности вероятности от параметра. Методы классификации, использующие лишь расстояние между классифицируемыми объектами, могут применяться к совокупностям объектов произвольной природы, лишь бы в пространстве, где они лежат, была задана метрика. Цель статистики объектов нечисловой природы состоит в том, чтобы систематически рассматривать методы статистической обработки данных как произвольной природы, так и представляющих собой указанные выше конкретные виды объектов нечисловой природы, т.е. методы описания данных, оцениванию и проверки гипотез. Взгляд с общей точки зрения позволяет получить новые результаты и в других областях математической статистики.



**Использование объектов нечисловой природы при формировании математической модели реального явления.**

Использование объектов нечисловой природы часто порождено желанием обрабатывать более объективную, более освобожденную от погрешностей информацию. "Как показали многочисленные опыты, человек более правильно ( и с меньшими затруднениями) отвечает на вопросы качественного например, сравнительного, характера, чем количественного. Так, ему легче сказать, какая из двух гирь тяжелее, чем указать их примерный вес в граммах" [87,с.3]. Другими словами, использование объектов нечисловой природы - средство повысить устойчивость математических моделей реальных явлений. Сначала конкретные области статистики объектов нечисловой природы (а именно, прикладная теория измерений, нечеткие и случайные множества) были рассмотрены как частные постановки проблемы устойчивости математических моделей реальных явлений к допустимым колебаниям исходных данных и предпосылок модели [37,гл.3,4], а затем была понята необходимость проведения работ по развитию статистики объектов нечисловой природы как самостоятельного научного направления [84].

Начнем со шкал измерения. "Науку о единстве мер и точности измерений называют метрологией" [88,с.5].Таким образом, репрезентативная теория измерений - часть метрологии [89]. "Методы обработки данных должны быть адекватны относительно допустимых преобразований шкал измерения в смысле репрезентативной теории измерений" [90 §4.1]. Однако установление типа шкалы, т.е. задания группы - дело специалиста соответствующей прикладной области. Так, оценки привлекательности профессий мы считали измеренными в порядковой шкале. Однако отдельные социологи не соглашались с этим, считая, что выпускники школ пользуются шкалой с более узкой группой допустимых преобразований, например, интервальной шкалой. Очевидно, эта проблема относится не к математике, а к наукам о человеке. Для ее решения может быть поставлен эксперимент (достаточно трудоемкий), описанный в работе [38]. Пока же он не поставлен, целесообразно принимать порядковую шкалу, так как это гарантирует от возможных ошибок.



"Другими известными примерами порядковых шкал являются: в медицине - шкала стадий гипертонической болезни по Мясникову, шкала степеней сердечной недостаточности по Стражеско-Василенко-Лангу, шкала степени выраженности коронарной недостаточности по Фогельсону; в минералогии - шкала Мооса (тальк - 1, гипс - 2, кальций - 3, флюорит - 4, апатит - 5, ортоклаз - 6, кварц - 7, топаз - 8, корунд - 9, алмаз - 10), по которому минералы классифицируются согласно критерию твердости; в географии - бофортова шкала ветров ("штиль", "слабый ветер", "умеренный ветер" и т.д.) [91, с. 329].

По шкале интервалов измеряют величину потенциальной энергии или координату точки на прямой, на которой не отмечены ни начало, ни единица измерения; по шкале отношений - большинство физических единиц: массу тела, длину, заряд, а также цены в экономике. Время измеряется по шкале разностей, если год принимаем естественной единицей измерения, и по шкале интервалов в общем случае. В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по порядковой шкале (холоднее - теплее), затем - по интервальной (шкалы Цельсия, Фаренгейта, Реомюра) и, наконец, после открытия абсолютного нуля температур - по шкале отношений (шкала Кельвина) [89]. Следует отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины [37, 39, 40, 63, 89]. Термин "репрезентативная" разъяснен в работах [37, 39, 40]. Он использовался, чтобы отличить рассматриваемый подход к измерениям от классической метрологии [88], от работ А.Н.Колмогорова и А.Лебега, связанных с измерением геометрических величин (например, [92])., от "алгоритмической теории измерения" [93] и др.

Необходимость использования в математических моделях реальных явлений таких объектов нечисловой природы, как бинарные отношения, множества, нечеткие множества, кратко была показана выше. Здесь же обратим внимание, что используемые в классической статистике результаты наблюдений также "не совсем числа". Именно любая величина измеряется всегда с некоторой погрешностью и результатом наблюдения является



Погрешностями измерений занимается метрология [88].Отметим справедливость следующих фактов:

а) для большинства реальных измерений невозможно полностью исключить систематическую ошибку: ;



б) распределение не всегда является нормальным [94];



в) и обычно нельзя считать независимыми случайными величинами;



г) распределение погрешностей оценивается по результатам случайных наблюдений, следовательно, полностью известным считать его нельзя; зачастую исследователь располагает лишь границами для систематической погрешности и оценками таких характеристик для случайной погрешности, как дисперсия или размах.

Приведенные факты показывают ограниченность области применимости модели погрешностей, в которой и рассматриваются как независимые случайные величины, причем имеет нормальное распределение с нулевым математическим ожиданием.



Строго говоря, результаты наблюдения всегда имеют дискретное распределение, поскольку описываются числами с небольшими (1 - 5) числом значащих цифр. Возникает дилемма: либо признать, что непрерывные распределения - фикция, и прекратить ими пользоваться, либо считать, что непрерывные распределения имеют "реальные" величины , которые мы наблюдаем с принципиально неустранимой погрешностью . Первый выход в настоящее время нецелесообразен, так как потребует отказаться от большей части разработанного математического аппарата. Из второго следует необходимость изучения влияния неустранимых погрешностей на статистические выводы.



Погрешности можно учитывать либо с помощью вероятностной модели (- случайная величина, имеющая функцию распределения, вообще говоря, зависящую от ), либо с помощью нечетких множеств. Во втором случае приходим к теории нечетких чисел, развитой П.Б.Шошиным [95] с целью описания поведения человека, и к интервальной статистике [9, 13, 19 - 25, 96 - 101].



Другой источник появления связан с принятой в конструкторской и технологической документации системой допусков на контролируемые параметры изделий и деталей, с использованием шаблонов при проверке контроля качества продукции. В этих случаях характеристики определяются не свойствами средств измерения, а применяемой технологией проектирования и производства. В терминах математической статистики сказанному соответствует группировка данных, при которой мы знаем, какому из заданных интервалов принадлежит наблюдение, но не знаем точного значения результата наблюдения. Применение группировки может дать экономический эффект, поскольку зачастую легче (в среднем) установить, к какому интервалу относится результат наблюдения, чем точно измерить его.



**Объекты нечисловой природы как результат статистической обработки данных.**

Объекты нечисловой природы появляются не только на "входе" статистической процедуры, но и в процессе обработки данных, и на "выходе" в качестве итога статистического анализа.

Рассмотрим простейшую прикладную постановку задачи регрессии. Данные имеют вид . Цель состоит в том, чтобы с достаточной точностью описать как полином от , т.е. модель имеет вид



, (5)



где - неизвестная степень полинома; - неизвестные коэффициенты многочлена; , - погрешности, которые для простоты примем независимыми и имеющими одно и то же нормальное распределение. Распространенная процедура такова [102]: сначала пытаются применить модель (5) для линейной функции ( = 1), при неудаче переходят к многочлену второго порядка ( = 2), если снова неудача, то берут модель (5) с = 3 и т.д. (адекватность модели проверяют по F-критерию Фишера).



Обсудим свойства этой процедуры в терминах математической статистики. Если степень полинома задана ( = ), то его коэффициенты оценивают методом наименьших квадратов, свойства этих оценок хорошо известны (см., например, [62, гл.26}). Однако в описанной выше реальной постановке тоже является неизвестным параметром и подлежит оценке. Таким образом, требуется оценить объект , ., множество значений которого можно обозначить Это - объект нечисловой природы, обычные методы оценивания его неприменимы, так как - дискретный параметр. В рассматриваемой постановке методы оценивания носят в основном эвристический характер {103, гл.12}. Свойства описанной выше распространенной процедуры рассмотрены в работе {104}; в которой показано, что m при этом оценивается несостоятельно (см. также .{14,18}).



В более общем случае линейной регрессии данные имеют вид , где - вектор предикторов (объясняющих переменных), а модель



(6)



( - некоторое подмножество множества ;- те же, что и в модели (5); - неизвестные коэффициенты при предикторах с номерами из {103]). Модель (5) сводится к модели (6), если



.,



В модели (5) есть естественный порядок ввода предикторов в рассмотрение - в соответствии с возрастанием степени, а в модели (6) естественного порядка нет, поэтому здесь стоит произвольное подмножество множества предикторов. Есть только частичный порядок - чем мощность подмножества меньше, тем лучше. Модель (6) особенно актуальна в задачах управления качеством продукции, в медицине и социологии, когда из большого числа факторов, предположительно влияющих на изучаемую переменную, надо отобрать по возможности наименьшее число значимых факторов и с их помощью сконструировать прогнозирующую формулу (6).

Задача оценивания модели (6) разбивается на две последовательные задачи: оценивание множества - подмножества множества всех предикторов, а затем - неизвестных параметров . Методы решения второй задачи хорошо известны и изучены. Гораздо хуже обстоит дело с оцениванием объекта нечисловой природы . Существующие методы [103] - в основном эвристические, они зачастую не являются даже состоятельными. Понятие состоятельности в данном случае требует определения. Пусть - истинное подмножество предикторов, т.е. подмножество, для которого справедлива модель (6), а подмножество предикторов - его оценка. Оценка является состоятельной, если



,



где - знак симметрической разности множеств; означает число элементов в множестве , а предел понимается в смысле сходимости по вероятности.



Задача оценивания в моделях регрессии, таким образом, разбивается на две - оценивание структуры модели и параметров при заданной структуре. в модели (5) структура описывается неотрицательным целым числом , в модели (6) - множеством . Структура - объект нечисловой природы. Задача ее оценивания сложна, в то время как задача оценивания численных параметров при заданной структуре хорошо изучена, разработаны эффективные ( в смысле математической статистики) методы.



Такова же ситуацию в факторном анализе (включая метод главных компонент) и многомерном шкалировании [38]. Ряд других примеров можно найти в списке оптимизационных постановок основных проблем прикладного многомерного статистического анализа [91].

Перейдем к объектам нечисловой природы на "выходе" статистической процедуры. Примеры многочисленны. Разбиения - итог работы многих алгоритмов классификации, в частности алгоритм кластер-анализа. Ранжировки - результат упорядочения профессий по привлекательности, автоматизированной обработки мнений экспертов - членов комиссии по подведению итогов конкурса научных работ [105] или итогов конкурса по решению задач в Вечерней математической школе [106]. (В двух последних случаях используются ранжировки со связями; так, в одну группу, наиболее многочисленную, попадают работы, не получившие наград.) Из всех объектов нечисловой природы, видимо, наиболее часты на "выходе" дихотомические данные - принять или не принять гипотезу, в частности принять или забраковать партию продукции [58]. Дихотомические данные используются научными исследованиями [46]. Результатом статистической обработка данных может быть множество, например зона наибольшего поражения [107], или последовательность множеств, например "среднемерное" описание распространения пожара [68]. Нечетким множеством Э.Борель [55] предлагал описывать представление людей о числе зерен, образующем "кучу". С помощью нечетких множеств формализуются значения лингвистических переменных, выступающих как итоговая оценка качества систем автоматизированного проектирования, сельскохозяйственных машин [108], бытовых газовых плит [109], надежности программного обеспечения [110, 111] или систем управления. Можно констатировать, что все виды объектов нечисловой природы могут появляться " на выходе" статистического исследования.

**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.