Математические методы исследования

УДК 519.2

СТАТИСТИКА ИНТЕРВАЛЬНЫХ ДАННЫХ (обобщающая статья)

© А. И. Орлов[1]

Статья поступила 11 апреля 2014 г.

Рассмотрены основные идеи асимптотической математической статистики интервальных данных, в которой элементы выборки — не числа, а интервалы. Алгоритмы и выводы статистики интервальных данных принципиально отличаются от классических. Приведены результаты, связанные с основополагающими понятиями нотны и рационального объема выборки. Статистика интервальных данных является составной частью статистики объектов нечисловой природы.

Ключевые слова: математическая статистика; интервальные данные; асимптотические методы; нотна; рациональный объем выборки; оценивание; статистика объектов нечисловой природы.

Статистика интервальных данных — область математической статистики, в которой элементы выборки — не числа, а интервалы. Это приводит к алгоритмам и выводам, принципиально отличающимся от классических, относящихся к числовым данным. Настоящая работа посвящена основным идеям и подходам асимптотической статистики интервальных данных. Впервые дается анализ развития этой области математической статистики — от первых работ до современных. Приведены результаты, связанные с основополагающими в рассматриваемой области прикладной математической статистики понятиями нотны и рационального объема выборки.

Развитие статистики интервальных данных

Перспективная и быстро развивающаяся область статистических исследований последних десятилетий — математическая статистика интервальных данных. Речь идет о развитии методов прикладной математической статистики в ситуации, когда статистические данные — не числа, а интервалы, в частности, порожденные наложением ошибок измерения на значения случайных величин [1, 2]. Приведем основные идеи весьма перспективного для вероятностно-статистических методов и моделей принятия решений асимптотического направления в статистике интервальных данных.

В настоящее время признается необходимым изучение устойчивости (робастности) оценок параметров к малым отклонениям исходных данных и предпосылок модели. Однако популярная среди теоретиков модель «засорения» (Тьюки – Хьюбера) представляется не вполне адекватной. Эта модель нацелена на изучение влияния больших «выбросов». Поскольку любые реальные измерения лежат в некотором фиксированном диапазоне, заданном в техническом паспорте средства измерения, то зачастую выбросы не могут быть слишком большими. Поэтому представляются полезными иные, более общие схемы устойчивости, введенные в монографии [3], в которых, например, учитываются отклонения распределений результатов наблюдений от предположений модели.

В одной из таких схем изучается влияние интервальности исходных данных на статистические выводы. Необходимость такого изучения стала очевидной следующим образом. В государственных стандартах СССР по прикладной статистике в обязательном порядке давалось справочное приложение «Примеры применения правил стандарта». При подготовке ГОСТ 11.011–83 [4] разработчикам стандарта были переданы для анализа реальные данные о наработке резцов до предельного состояния (в часах). Оказалось, что все эти данные представляли собой либо целые числа, либо полуцелые (т.е. после умножения на два становящиеся целыми). Ясно, что исходная длительность наработок искажена. Необходимо было учесть в статистических процедурах наличие такого искажения исходных данных. Как это сделать?

Первое, что могло показаться верным, — модель группировки данных, согласно которой для истинного значения X проводится замена на ближайшее число из множества $\{0{,}5n,\ n = 1, 2, 3, ...\}$. Однако эту модель целесообразно подвергнуть сомнению, а также рассмотреть иные модели. Возможно, что X надо приво-

[1] Институт высоких статистических технологий и эконометрики Московского государственного технического университета им. Н. Э. Баумана, Москва, Россия; Московский физико-технический институт, Москва, Россия; Центральный научно-исследовательский институт машиностроения, г. Королёв Московской области, Россия; e-mail: prof-orlov@mail.ru

дить к ближайшему сверху элементу указанного множества — в случае, если проверка качества поставленных на испытание резцов проводилась раз в полчаса. Другой вариант: если расстояния от X до двух ближайших элементов множества $\{0{,}5n,\ n = 1, 2, 3, ...\}$ примерно равны, то естественно ввести рандомизацию при выборе заменяющего числа, и т.д.

Целесообразно построить новую математико-статистическую модель, согласно которой результаты наблюдений — не числа, а интервалы. Например, если в таблице приведено значение 53,5, то это значит, что реальное значение — какое-то число от 53,0 до 54,0, т.е. какое-то число в интервале [53,5 – 0,5; 53,5 + 0,5], где 0,5 — максимально возможная погрешность. Принимая эту модель, мы попадаем в новую научную область — статистику интервальных данных [5, 6]. Статистика интервальных данных идейно связана с интервальной математикой, в которой в роли чисел выступают интервалы (см., например, монографию [7]). 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.

В интервальной математике сумма двух интервальных чисел $[a, b]$ и $[c, d]$ имеет вид $[a, b] + [c, d] = [a + c, b + d]$, а разность находится по формуле $[a, b] - [c, d] = [a - d, b - c]$. Для положительных a, b, c, d произведение определяется формулой $[a, b] \cdot [c, d] = [ac, bd]$, а частное имеет вид $[a, b]/[c, d] = [a/d, b/c]$. Эти формулы получены при решении соответствующих оптимизационных задач. Пусть x лежит в отрезке $[a, b]$, а y — в отрезке $[c, d]$. Каковы минимальное и максимальное значения для $x + y$? Очевидно, $a + c$ и $b + d$ соответственно. Минимальные и максимальные значения для $x - y$, xy, x/y указывают нижние и верхние границы для интервальных чисел, задающих результаты арифметических операций. А от арифметических операций можно перейти ко всем остальным математическим алгоритмам. Так строится интервальная математика.

Как известно (в частности, из работы [2]), исследователям удалось решить ряд задач теории интервальных дифференциальных уравнений, в которых коэффициенты, начальные условия и решения описываются с помощью интервалов. По мнению ряда специалистов, статистика интервальных данных является частью интервальной математики [7]. Впрочем, распространена и другая точка зрения, согласно которой такое включение нецелесообразно, поскольку статистика интервальных данных использует несколько иные подходы к алгоритмам анализа реальных данных, чем сложившиеся в интервальной математике.

Рассмотрим 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 классической математической статистике предельные переходы осуществляются в обратном порядке — сначала уменьшаются до нуля погрешности измерений, и только затем устремляется к бесконечности объем выборки). В частности, еще в начале 1980-х годов с помощью такой асимптотики сформулированы правила выбора метода оценивания в ГОСТ 11.011–83 [4].

Нами разработана [8] 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 и дисперсии [1], медианы и коэффициента вариации [9], параметров гамма-распределения [4, 10] и характеристик аддитивных статистик [8], при проверке гипотез о параметрах нормального распределения, в том числе с помощью критерия Стьюдента, а также гипотезы однородности с помощью критерия Смирнова [9]. 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 [11].

Разработаны подходы к рассмотрению интервальных данных в основных постановках регрессионного, дискриминантного и кластерного анализов [12]. 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 [12, 13]. Проведена первоначальная разработка интервального дискриминантного анализа, рассмотрено влияние интервальности данных на показатель качества классификации [12, 14]. Основные идеи и результаты указанного направления в статистике интервальных данных приведены в публикациях обзорного характера [5, 6].

Как показала Международная конференция ИНТЕРВАЛ-92 [2], в области асимптотической математической статистики интервальных данных мы имеем мировой приоритет. По нашему мнению, 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.

Многие из утверждений статистики интервальных данных весьма отличаются от аналогов из классической математической статистики. В частности, не существует состоятельных оценок; средний квадрат ошибки оценки, как правило, асимптотически равен сумме дисперсии оценки, рассчитанной согласно классической теории, и некоторого положительного числа, равного квадрату так называемой нотны — максимально возможного отклонения значения статистики из-за погрешностей исходных данных (в результате метод моментов оказывается иногда точнее метода максимального правдоподобия [11]); нецелесообразно увеличивать объем выборки сверх некоторого предела, называемого рациональным объемом выборки (вопреки классической теории, согласно которой чем больше объем выборки, тем точнее выводы).

В стандарт [4] включен раздел 5, посвященный выбору метода оценивания при неизвестных параметрах формы и масштаба и известном параметре сдвига и основанный на концепциях статистики интервальных данных. Теоретическое обоснование этого раздела стандарта было опубликовано лишь через пять лет в статье [10].

В 1982 г. при разработке стандарта [4] сформулированы основные идеи статистики интервальных данных. Однако из-за недостатка времени они не были полностью реализованы в ГОСТ 11.011–83, он написан в основном в классической манере. Развитие идей статистики интервальных данных продолжается уже более 30 лет, их большое значение для современной прикладной статистики обосновано в [15, 16].

Вторая ведущая научная школа в области статистики интервальных данных — это школа проф. А. П. Вощинина (1937 – 2008 гг.), активно работающая с конца 70-х годов. Полученные результаты отражены в ряде монографий (см., прежде всего, [17 – 19]), статей [1, 20, 21], докладов [2], диссертациях [22, 23]. Изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности.

Наше научное направление отличается нацеленностью на асимптотические результаты, полученные при больших объемах выборок и малых погрешностях измерений, поэтому его полное название таково: асимптотическая математическая статистика интервальных данных.

Основные идеи статистики интервальных данных

Сформулируем сначала основные идеи асимптотической математической статистики интервальных данных, а затем рассмотрим реализацию этих идей на перечисленных выше примерах. Основные идеи достаточно просты, в то время как их проработка в конкретных ситуациях зачастую оказывается достаточно трудоемкой.

Пусть существо реального явления описывается выборкой $x_1, x_2, ..., x_n$. В вероятностной теории математической статистики, из которой мы исходим [24], выборка — это набор независимых в совокупности одинаково распределенных случайных величин. Однако беспристрастный и тщательный анализ подавляющего большинства реальных задач показывает, что статистику известна отнюдь не выборка $x_1, x_2, ..., x_n$, а величины

$$y_j = x_j + \varepsilon_j,\ j = 1, 2, ..., n,$$

где $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ — некоторые погрешности измерений, наблюдений, анализов, опытов, исследований (например, инструментальные ошибки).

Одна из причин появления погрешностей — запись результатов наблюдений с конечным числом значащих цифр. Дело в том, что для случайных величин с непрерывными функциями распределения событие, состоящее в попадании хотя бы одного элемента выборки в множество рациональных чисел, согласно правилам теории вероятностей имеет вероятность 0, а такими событиями в теории вероятностей принято пренебрегать. Поэтому при рассуждениях о выборках из нормального, логарифмически нормального, экспоненциального, равномерного, гамма-распределений, распределения Вейбулла – Гнеденко и др. приходится принимать, что эти распределения имеют элементы исходной выборки $x_1, x_2, ..., x_n$, в то время как статистической обработке доступны лишь искаженные значения $y_j = x_j + \varepsilon_j$.

Введем обозначения:

$$\mathbf{x} = (x_1, x_2, ..., x_n),\ \mathbf{y} = (y_1, y_2, ..., y_n),$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n).$$

Пусть статистические выводы основываются на статистике f: $R^n \to R^1$, используемой для оценивания параметров и характеристик распределения, проверки гипотез и решения иных статистических задач. Принципиально важная для статистики интервальных данных идея такова: *статистик знает только $f(\mathbf{y})$, но не $f(\mathbf{x})$.*

Очевидно, в статистических выводах необходимо отразить различие между $f(\mathbf{y})$ и $f(\mathbf{x})$. Одним из двух основных понятий статистики интервальных данных является понятие нотны.

Определение. Величину максимально возможного (по абсолютной величине) отклонения, вызванного

погрешностями наблюдений **ε**, известного статистику значения $f(\mathbf{y})$ от истинного значения $f(\mathbf{x})$, т.е.

$$N_f(\mathbf{x}) = \sup|f(\mathbf{y}) - f(\mathbf{x})|,$$

где супремум берется по множеству возможных значений вектора погрешностей **ε** (см. ниже), будем называть *нотной*.

Если функция f имеет частные производные второго порядка, а ограничения на погрешности

$$|\varepsilon_i| \le \Delta, \ \ i = 1, 2, ..., n, \qquad (1)$$

причем Δ мало, то приращение функции f с точностью до бесконечно малых более высокого порядка описывается главным линейным членом, т.е.

$$f(\mathbf{y}) - f(\mathbf{x}) = \sum_{1\le i\le n} \frac{\partial f(\mathbf{x})}{\partial x_i}\varepsilon_i + O(\Delta^2).$$

Чтобы получить асимптотическое (при $\Delta \to 0$) выражение для нотны, достаточно найти максимум и минимум линейной функции (главного линейного члена) на кубе, заданном неравенствами (1). Совершенно очевидно, что максимум достигается, если

$$\varepsilon_i = \begin{cases} \Delta, \dfrac{\partial f(\mathbf{x})}{\partial x_i} \ge 0, \\ -\Delta, \dfrac{\partial f(\mathbf{x})}{\partial x_i} < 0, \end{cases}$$

а минимум, отличающийся от максимума только знаком, — при $\varepsilon_i = -\varepsilon_i$. Следовательно, *нотна* с точностью до бесконечно малых более высокого порядка имеет вид

$$N_f(\mathbf{x}) = \Delta\left(\sum_{1\le i\le n}\left|\frac{\partial f(\mathbf{x})}{\partial x_i}\right|\right).$$

Это выражение назовем *асимптотической нотной*.

Условие (1) означает, что исходные данные представляются статистику в виде интервалов $[y_i - \Delta; y_i + \Delta]$, $i = 1, 2, ..., n$ (отсюда и название этого научного направления). Ограничения на погрешности могут задаваться разными способами — кроме абсолютных ошибок используются относительные или иные показатели различия между **x** и **y**.

Если задана не предельная абсолютная погрешность Δ, а предельная относительная погрешность δ, т.е. ограничения на погрешности вошедших в выборку результатов измерений имеют вид

$$|\varepsilon_i| \le \delta|x_i|, \ \ i = 1, 2, ..., n,$$

то аналогичным образом получаем, что нотна с точностью до бесконечно малых более высокого порядка, т.е. асимптотическая нотна, имеет вид

$$N_f(\mathbf{x}) = \delta\left(\sum_{1\le i\le n}\left|x_i\frac{\partial f(\mathbf{x})}{\partial x_i}\right|\right).$$

При практическом использовании рассматриваемой концепции необходимо провести тотальную замену символов x на символы y. В каждом конкретном случае удается показать, что в силу малости погрешностей разность $N_f(\mathbf{y}) - N_f(\mathbf{x})$ является бесконечно малой более высокого порядка, чем $N_f(\mathbf{x})$ или $N_f(\mathbf{y})$.

Основные результаты в вероятностной модели

В классической вероятностной модели элементы исходной выборки $x_1, x_2, ..., x_n$ рассматриваются как независимые одинаково распределенные случайные величины. Как правило, существует некоторая константа $C > 0$ такая, что в смысле сходимости по вероятности

$$\lim_{n\to\infty} N_f(\mathbf{x}) = C\Delta. \qquad (2)$$

Соотношение (2) доказывается отдельно для каждой конкретной задачи.

При использовании классических статистических методов в большинстве случаев статистика $f(\mathbf{x})$ является асимптотически нормальной. Это означает, что существуют константы a и σ^2 такие, что

$$\lim_{n\to\infty} P\left(\sqrt{n}\,\frac{f(\mathbf{x}) - a}{\sigma} < x\right) = \Phi(\mathbf{x}),$$

где $\Phi(\mathbf{x})$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. При этом обычно оказывается, что

$$\lim_{n\to\infty}\sqrt{n}(Mf(\mathbf{x}) - a) = 0 \ \text{ и } \ \lim_{n\to\infty} nDf(\mathbf{x}) = \sigma^2,$$

а потому в классической математической статистике средний квадрат ошибки статистической оценки

$$M[f(\mathbf{x}) - a]^2 = [Mf(\mathbf{x}) - a]^2 + Df(\mathbf{x}) = \sigma^2/n$$

с точностью до членов более высокого порядка.

В статистике интервальных данных ситуация совсем иная — обычно можно доказать, что средний квадрат ошибки

$$\max_{\{\varepsilon\}} M[f(\mathbf{x}) - a]^2 = \frac{\sigma^2}{n} + N_f^2(\mathbf{y}) + o\left(\Delta^2 + \frac{1}{n}\right). \qquad (3)$$

Из соотношения (3) вытекает ряд важных следствий. Правая часть этого равенства, в отличие от правой части соответствующего классического равенства, не стремится к нулю при безграничном возрастании объема выборки. Она остается больше некоторого

положительного числа, а именно, квадрата нотны. Следовательно, статистика $f(\mathbf{x})$ не является состоятельной оценкой параметра a. Более того, состоятельных оценок вообще не существует.

Пусть доверительным интервалом для параметра a, соответствующим заданной доверительной вероятности γ, в классической математической статистике является интервал $(c_n(\gamma); d_n(\gamma))$. В статистике интервальных данных аналогичный доверительный интервал является более широким. Он имеет вид $(c_n(\gamma) - N_f(\mathbf{y}); d_n(\gamma) + N_f(\mathbf{y}))$. Таким образом, его длина увеличивается на две нотны. Следовательно, при увеличении объема выборки длина доверительного интервала не может стать меньше, чем $2C\Delta$ (см. формулу (2)).

В статистике интервальных данных методы оценивания параметров имеют другие свойства по сравнению с классической математической статистикой. Так, при больших объемах выборок метод моментов может быть заметно лучше, чем метод максимального правдоподобия (т.е. иметь меньший средний квадрат ошибки (3)), в то время как в классической математической статистике второй из названных методов всегда не хуже первого.

Рациональный объем выборки

Анализ формулы (3) показывает, что в отличие от классической математической статистики нецелесообразно безгранично увеличивать объем выборки, поскольку средний квадрат ошибки остается всегда большим квадрата нотны. Поэтому представляется полезным ввести понятие «рационального объема выборки» n_{rat}, при достижении которого продолжать наблюдения нецелесообразно.

Как установить «рациональный объем выборки»? Можно воспользоваться идеей «принципа уравнивания погрешностей», предложенной в монографии [3]. Речь идет о том, что вклад погрешностей различной природы в общую погрешность должен быть примерно одинаков. Этот принцип дает возможность выбирать необходимую точность оценивания тех или иных характеристик в тех случаях, когда это зависит от исследователя. В статистике интервальных данных в соответствии с «принципом уравнивания погрешностей» предлагается определять рациональный объем выборки n_{rat} из условия равенства двух величин — метрологической составляющей, связанной с нотной, и статистической составляющей — в среднем квадрате ошибки (3), т.е. из условия

$$\frac{\sigma^2}{n_{rat}} = N_f^2(\mathbf{y}),\ n_{rat} = \frac{\sigma^2}{N_f^2}(\mathbf{y}).$$

Для практического использования выражения для рационального объема выборки неизвестные теоретические характеристики необходимо заменить их оценками. Это делается в каждой конкретной задаче по-своему.

Исследовательскую программу в области статистики интервальных данных кратко можно сформулировать так: для любого алгоритма анализа данных (алгоритма прикладной статистики) необходимо вычислить нотну и рациональный объем выборки. Можно вычислить иные величины из того же понятийного ряда, возникающие в многомерном случае, при наличии нескольких выборок и при иных обобщениях описываемой здесь простейшей схемы. Затем следует проследить влияние погрешностей исходных данных на точность оценивания, доверительные интервалы, значения статистик критериев при проверке гипотез, уровни значимости и другие характеристики статистических выводов. Очевидно, классическая математическая статистика является частью статистики интервальных данных, выделяемой условием $\Delta = 0$.

Поясним теоретические концепции статистики интервальных данных на простых примерах.

Оценивание математического ожидания

Пусть необходимо оценить математическое ожидание случайной величины с помощью обычной оценки — среднего арифметического результатов наблюдений, т.е.

$$f(\mathbf{x}) = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Тогда при справедливости ограничений (1) на абсолютные погрешности имеем $N_f(\mathbf{x}) = \Delta$. Таким образом, нотна полностью известна и не зависит от многомерной точки, в которой берется. Вполне естественно, что если каждый результат наблюдения известен с точностью до Δ, то и среднее арифметическое известно с той же точностью. Ведь возможна систематическая ошибка: если к каждому результату наблюдения добавить Δ, то и среднее арифметическое увеличится на Δ.

Поскольку

$$D(\bar{x}) = \frac{D(x_1)}{n},$$

то в ранее введенных обозначениях

$$\sigma^2 = D(x_1).$$

Следовательно, рациональный объем выборки

$$n_{rat} = D(x_1)/\Delta^2.$$

Для практического использования полученной формулы надо оценить дисперсию результатов наблюдений. Поскольку Δ мало, это можно сделать обыч-

ным способом, например, с помощью несмещенной выборочной оценки дисперсии

$$s^2(y) = \frac{1}{n-1} \sum_{1 \le i \le n} (y_i - \bar{y})^2.$$

Здесь и далее рассуждения часто ведутся на двух уровнях. Первый — это уровень «истинных» случайных величин, обозначаемых «x», описывающих реальность, но неизвестных специалисту по анализу данных. Второй — уровень известных этому специалисту величин «y», отличающихся погрешностями от истинных. Погрешности малы, поэтому функции от x отличаются от функций от y на некоторые бесконечно малые величины. Эти соображения и позволяют использовать $s^2(y)$ как оценку $D(x_1)$.

Итак, выборочной оценкой рационального объема выборки является

$$n_{sample\text{-}rat} = s^2(y)/\Delta^2.$$

Уже на этом первом рассматриваемом примере видим, что рациональный объем выборки находится не где-то вдали, а непосредственно рядом с теми объемами, с которыми имеет дело любой практически работающий статистик. Например, если статистик знает, что

$$\Delta = \sigma/6,$$

то $n_{rat} = 36$. Именно такова погрешность контрольных шаблонов во многих технологических процессах! Поэтому, занимаясь управлением качеством, необходимо обращать внимание на действующую на предприятии систему измерений.

По сравнению с классической математической статистикой доверительный интервал для математического ожидания (для заданной доверительной вероятности γ) имеет другой вид

$$\left(\bar{y} - \Delta - u(\gamma)\frac{s}{\sqrt{n}};\ \bar{y} + \Delta + u(\gamma)\frac{s}{\sqrt{n}} \right), \qquad (4)$$

где $u(\gamma)$ — квантиль порядка $(1+\gamma)/2$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

По поводу формулы (4) ведется продолжительная дискуссия. Отмечается, что она получена на основе Центральной предельной теоремы теории вероятностей и может быть использована при любом распределении результатов наблюдений (с конечной дисперсией). Если же имеется дополнительная информация, то, по мнению отдельных специалистов, формула (4) может быть уточнена. Например, если известно, что распределение x_i является нормальным, то в качестве $u(\gamma)$ целесообразно использовать квантиль распределения Стьюдента. К этому надо добавить, что по небольшому числу наблюдений нельзя надежно установить нормальность, а при росте объема выборки квантили распределения Стьюдента приближаются к квантилям нормального распределения.

Вопрос о том, часто ли результаты наблюдений имеют нормальное распределение, подробно обсуждался. Выяснилось, что распределения встречающихся в практических задачах результатов измерений почти всегда отличны от нормальных [25], а также от распределений из иных параметрических семейств, описываемых в учебниках.

Применительно к оцениванию математического ожидания (но не к оцениванию других характеристик или параметров распределения) факт существования границы возможной точности, определяемой точностью исходных данных, неоднократно отмечался в литературе ([26, с. 230 – 234], [27, с. 121] и др.).

Оценивание дисперсии

Для статистики $f(y) = s^2(y)$, где $s^2(y)$ — выборочная дисперсия (несмещенная оценка теоретической дисперсии), при справедливости ограничений (1) на абсолютные погрешности имеем

$$N_f(y) = \frac{2\Delta}{n-1} \sum_{i=1}^{n} |y_i - \bar{y}| + O(\Delta^2).$$

Можно показать, что нотна $N_f(y)$ сходится к

$$2\Delta M|x_1 - M(x_1)|$$

по вероятности с точностью до $O(\Delta)$, когда n стремится к бесконечности. Это же предельное соотношение верно и для нотны $N_f(x)$, вычисленной для исходных данных. Таким образом, в данном случае справедлива формула (2) с

$$C = 2M|x_1 - M(x_1)|.$$

Известно [28], что случайная величина

$$\frac{s^2 - \sigma^2}{\sqrt{n}}$$

является асимптотически нормальной с математическим ожиданием 0 и дисперсией $D(x_1^2)$.

Из сказанного следует, что в статистике интервальных данных асимптотический доверительный интервал для дисперсии σ^2 (соответствующий доверительной вероятности γ) имеет вид

$$(s^2(y) - A;\ s^2 + A),$$

где

$$A = \frac{u(\gamma)}{\sqrt{n(n-1)}} \sqrt{\sum_{i=1}^{n} \left(y_i^2 - \frac{1}{n} \sum_{j=1}^{n} y_j^2 \right)^2} + \frac{2\Delta}{n-1} \sum_{i=1}^{n} |y_i - \bar{y}|.$$

Здесь $u(\gamma)$ обозначает тот же самый квантиль стандартного нормального распределения, что и выше в случае оценивания математического ожидания.

При оценивании дисперсии рациональный объем выборки

$$n_{rat} = \frac{D(x_1^2)}{4\Delta^2 (M|x_1 - M(x_1)|)^2},$$

а выборочную оценку рационального объема выборки $n_{sample\text{-}rat}$ можно вычислить, заменяя теоретические моменты на соответствующие выборочные и используя доступные статистику результаты наблюдений, содержащие погрешности.

Что можно сказать о численной величине рационального объема выборки? Как и в случае оценивания математического ожидания, она отнюдь не выходит за пределы обычно используемых объемов выборок. Так, если распределение результатов наблюдений x_i является нормальным с математическим ожиданием 0 и дисперсией σ^2, то в результате вычисления моментов случайных величин в предыдущей формуле получаем

$$n_{rat} = \frac{\sigma^2}{\pi\Delta^2}.$$

Например, если $\Delta = \sigma/6$, то $n_{rat} = 11$. Это меньше, чем при оценивании математического ожидания в предыдущем примере.

Статистика интервальных данных в прикладной статистике

Кратко рассмотрим положение статистики интервальных данных (СИД) среди других методов описания неопределенностей и анализа данных.

Нечеткость и СИД. С формальной точки зрения описание нечеткости интервалом — это частный случай описания ее нечетким множеством. В СИД функция принадлежности нечеткого множества имеет специфический вид — она равна единице в некотором интервале и нулю вне его. Такая функция принадлежности описывается всего двумя параметрами (границами интервала). Эта простота описания делает математический аппарат СИД гораздо более прозрачным, чем аппарат теории нечеткости в общем случае. Это, в свою очередь, позволяет исследователю продвинуться дальше, чем при использовании функций принадлежности произвольного вида.

Интервальная математика и СИД. Можно было бы сказать, что СИД — часть интервальной математики, что она так соотносится с прикладной математической статистикой, как интервальная математика — с математикой в целом. Однако исторически сложилось так, что интервальная математика занимается прежде всего вычислительными погрешностями. С точки зрения интервальной математики две известные формулы для выборочной дисперсии, а именно

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2,$$

имеют разные погрешности. А с точки зрения СИД эти две формулы задают одну и ту же функцию, поэтому им соответствуют совпадающие нотны и рациональные объемы выборок. Интервальная математика прослеживает процесс вычислений, СИД этим не занимается. Необходимо отметить, что типовые постановки СИД могут быть перенесены в другие области математики и, наоборот; вычислительные алгоритмы прикладной математической статистики и СИД заслуживают изучения. Однако и то, и другое, скорее всего, дело будущего. Из уже известного отметим применение методов СИД при анализе такой характеристики финансовых потоков, как *NPV* — чистая текущая стоимость [29, гл. 9].

Математическая статистика и СИД. Математическая статистика и СИД отличаются тем, в каком порядке делаются предельные переходы $n \to \infty$ и $\Delta \to 0$. При этом СИД переходит в математическую статистику при $\Delta = 0$. Правда, тогда исчезают основные особенности СИД: нотна становится равной нулю, а рациональный объем выборки — бесконечности. Рассмотренные выше методы СИД разработаны в предположении, что погрешности малы (но не исчезают), а объем выборки велик. СИД расширяет классическую математическую статистику тем, что в исходных статистических данных каждое число заменяет интервалом. С другой стороны, можно считать СИД новым этапом развития математической статистики.

Статистика объектов нечисловой природы и СИД. Статистика объектов нечисловой природы (СОНП) [30] расширяет область применения классической математической статистики путем включения в нее новых видов статистических данных. Естественно, при этом появляются новые виды алгоритмов анализа статистических данных и новый математический аппарат (в частности, происходит переход от методов суммирования к методам оптимизации). С точки зрения СОНП частному виду новых статистических данных — интервальным данным — соответствует СИД. Напомним, что одно из двух основных понятий СИД — нотна — определяется как решение оптимизационной задачи. Однако СИД, изучая классические методы прикладной статистики применительно к интервальным данным, по математическому аппарату ближе к классической математической статистике, чем другие части СОНП, например, статистика бинарных отношений.

Робастные методы статистики и СИД. Если понимать робастность как теорию устойчивости статистических методов по отношению к допустимым отклонениям исходных данных и предпосылок модели [3], то в СИД рассматривается одна из естественных постановок робастности. Однако в массовом сознании специалистов термин «робастность» закрепился за моделью засорения выборки большими выбросами (модель Тьюки – Хубера), хотя эта модель не имеет большого практического значения [31]. К этой модели СИД не имеет отношения.

Теория устойчивости и СИД. Общей схеме устойчивости [3, 32, 33] математических моделей социально-экономических явлений и процессов по отношению к допустимым отклонениям исходных данных и предпосылок моделей СИД полностью соответствует. Она посвящена математико-статистическим моделям, используемым при анализе статистических данных, а допустимые отклонения — это интервалы, заданные ограничениями на погрешности. СИД можно рассматривать как пример теории, в которой учет устойчивости позволил сделать нетривиальные выводы. Отметим, что с точки зрения общей схемы [3] устойчивость по Ляпунову в теории дифференциальных уравнений — весьма частный случай, в котором из-за его конкретности удалось весьма далеко продвинуться.

Минимаксные методы, типовые отклонения и СИД. Постановки СИД относятся к минимаксным. За основу берется максимально возможное отклонение. Это — «подход пессимиста», применяемый, например, в теории антагонистических игр. Использование минимаксного подхода позволяет подозревать СИД в завышении роли погрешностей измерения. Однако примеры изучения вероятностно-статистических моделей погрешностей, проведенные, в частности, при разработке методов оценивания параметров гамма-распределения [4, 10], показали, что это подозрение не подтверждается. Влияние погрешностей измерений по порядку такое же, только вместо максимально возможного отклонения (нотны) приходится рассматривать математическое ожидание соответствующего отклонения. Подчеркнем, что применение в СИД вероятностно-статистических моделей погрешностей не менее перспективно, чем минимаксных.

Научная школа А. П. Вощинина и СИД. Если в математической статистике неопределенность только статистическая, то в научной школе А. П. Вощинина — только интервальная. Можно сказать, что СИД лежит между классической прикладной математической статистикой и областью исследований научной школы А. П. Вощинина. Другое отличие состоит в том, что в этой школе разрабатывают новые методы анализа интервальных данных, а в СИД в настоящее время изучают устойчивость классических статистических методов по отношению к малым погрешностям. Подход СИД оправдывается распространенностью этих методов, однако в дальнейшем следует переходить к разработке новых методов, специально предназначенных для анализа интервальных данных.

Анализ чувствительности и СИД. При анализе чувствительности, как и в СИД, рассчитывают производные по используемым переменным или непосредственно находят изменения при отклонении переменной на ±10 % от базового значения. Однако этот анализ делают по каждой переменной отдельно. В СИД все переменные рассматривают совместно и находят максимально возможное отклонение (нотну). При малых погрешностях удается на основе главного члена разложения функции в многомерный ряд Тейлора получить удобную формулу для нотны. Можно сказать, что СИД — это многомерный анализ чувствительности.

Асимптотической математической статистике интервальных данных посвящены главы в учебниках [31, 34, 35]. Развиваются научные исследования как в научной школе А. П. Вощинина [36, 37], так и в СИД [38, 39].

По нашему мнению, во все виды статистического программного обеспечения должны быть включены алгоритмы интервальной статистики, «параллельные» обычно используемым в настоящее время алгоритмам прикладной математической статистики. Это позволит в явном виде учесть наличие погрешностей у результатов наблюдений (измерений, испытаний, анализов, опытов).

ЛИТЕРАТУРА

1. Дискуссия по анализу интервальных данных / Заводская лаборатория. 1990. Т. 56. № 7. С. 75 – 95.
2. Сборник трудов Международной конференции по интервальным и стохастическим методам в науке и технике (ИНТЕРВАЛ-92). — М.: МЭИ, 1992. Т. 1. — 216 с.; Т. 2. — 152 с.
3. **Орлов А. И.** Устойчивость в социально-экономических моделях. — М.: Наука, 1979. — 296 с.
4. ГОСТ 11.011–83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения. — М.: Изд-во стандартов, 1984. — 53 с.
5. **Orlov A. I.** Interval statistics / Interval Computations, 1992, № 1(3). P. 44 – 52.
6. **Орлов А. И.** Основные идеи интервальной математической статистики / Наука и технология в России. 1994. № 4(6). С. 8 – 9.
7. **Шокин Ю. И.** Интервальный анализ. — Новосибирск: Наука, 1981. — 112 с.
8. **Орлов А. И.** О развитии реалистической статистики. — В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. — Пермь: Изд-во Пермского государственного университета, 1990. С. 89 – 99.
9. **Орлов А. И.** Некоторые алгоритмы реалистической статистики. — В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. — Пермь: Изд-во Пермского государственного университета, 1991. С. 77 – 86.
10. **Орлов А. И.** О влиянии погрешностей наблюдений на свойства статистических процедур (на примере гамма-распределения). — В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. — Пермь: Изд-во Пермского государственного университета, 1988. С. 45 – 55.
11. **Орлов А. И.** Интервальная статистика: метод максимального правдоподобия и метод моментов. — В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. — Пермь: Изд-во Пермского государственного университета, 1995. С. 114 – 124.
12. **Орлов А. И.** Интервальный статистический анализ. — В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. — Пермь: Пермский государственный университет, 1993. С. 149 – 158.

13. **Биттар А. Б.** Метод наименьших квадратов для интервальных данных. Дипломная работа. — М.: МЭИ, 1994. — 38 с.
14. **Пузикова Д. А.** Об интервальных методах статистической классификации / Наука и технология в России. 1995. № 2(8). С. 12 – 13.
15. **Орлов А. И.** Пути развития статистических методов: непараметрика, робастность, бутстреп и реалистическая статистика / Надежность и контроль качества. 1991. № 8. С. 3 – 8.
16. **Орлов А. И.** Современная прикладная статистика / Заводская лаборатория. Диагностика материалов. 1998. Т. 64. № 3. С. 52 – 60.
17. **Вощинин А. П.** Метод оптимизации объектов по интервальным моделям целевой функции. — М.: МЭИ, 1987. — 109 с.
18. **Вощинин А. П., Сотиров Г. Р.** Оптимизация в условиях неопределенности. — М.: МЭИ; София: Техника, 1989. — 224 с.
19. **Вощинин А. П., Акматбеков Р. А.** Оптимизация по регрессионным моделям и планирование эксперимента. — Бишкек: Илим, 1991. — 164 с.
20. **Вощинин А. П.** Метод анализа данных с интервальными ошибками в задачах проверки гипотез и оценивания параметров неявных и линейно параметризованных функций / Заводская лаборатория. Диагностика материалов. 2000. Т. 66. № 3. С. 51 – 65.
21. **Вощинин А. П.** Интервальный анализ данных: развитие и перспективы / Заводская лаборатория. Диагностика материалов. 2002. Т. 68. № 1. С. 118 – 126.
22. **Дывак Н. П.** Разработка методов оптимального планирования эксперимента и анализа интервальных данных. Автореф. дис. ... канд. техн. наук. — М., 1992. — 20 с.
23. **Симов С. Ж.** Разработка и исследование интервальных моделей при анализе данных и проектировании экспертных систем. Автореф. дис. ... канд. техн. наук. — М., 1992. — 20 с.
24. **Орлов А. И.** Вероятность и прикладная статистика: основные факты: справочник. — М.: КНОРУС, 2010. — 192 с.
25. **Орлов А. И.** Часто ли распределение результатов наблюдений является нормальным? / Заводская лаборатория. 1991. Т. 57. № 7. С. 64 – 66.
26. **Новицкий П. В., Зограф И. А.** Оценка погрешностей результатов измерений. — Л.: Энергоатомиздат, 1985. — 248 с.
27. **Гнеденко Б. В., Хинчин А. Я.** Элементарное введение в теорию вероятностей. — М.: Наука, 1970. — 128 с.
28. **Боровков А. А.** Математическая статистика. — М.: Наука, 1984. — 472 с.
29. **Орлов А. И.** Эконометрика. Изд. 3-е, испр. и доп. — М.: Экзамен, 2004. — 576 с.
30. **Орлов А. И.** Тридцать лет статистики объектов нечисловой природы (обзор) / Заводская лаборатория. Диагностика материалов. 2009. Т. 75. № 5. С. 55 – 64.
31. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
32. **Орлов А. И.** Устойчивые экономико-математические методы и модели. — Saarbrücken: Lambert Academic Publishing, 2011. — 436 с.
33. **Орлов А. И.** Устойчивые математические методы и модели / Заводская лаборатория. Диагностика материалов. 2010. Т. 76. № 3. С. 59 – 67.
34. **Орлов А. И.** Теория принятия решений. — М.: Экзамен, 2006. — 574 с.
35. **Орлов А. И.** Организационно-экономическое моделирование: учебник. В 3 ч. Ч. 1. Нечисловая статистика. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2009. — 541 с.
36. **Вощинин А. П., Бронз П. В.** Построение аналитических моделей по данным вычислительного эксперимента в задачах анализа чувствительности и оценки экономических рисков / Заводская лаборатория. Диагностика материалов. 2007. Т. 73. № 1. С. 101 – 109.
37. **Вощинин А. П., Скибицкий Н. В.** Интервальный подход к выражению неопределенности измерений и калибровке цифровых измерительных систем / Заводская лаборатория. Диагностика материалов. 2007. Т. 73. № 11. С. 66 – 71.
38. **Орлов А. И.** Об оценивании параметров гамма-распределения / Журнал «Обозрение прикладной и промышленной математики». 1997. Т. 4. Вып. 3. С. 471 – 482.
39. **Гуськова Е. А., Орлов А. И.** Интервальная линейная парная регрессия / Заводская лаборатория. Диагностика материалов. 2005. Т. 71. № 3. С. 57 – 63.