

## **2.3. Статистика интервальных данных**

В статистике интервальных данных элементы выборки - не числа, а интервалы. Это приводит к алгоритмам и выводам, принципиально отличающимся от классических. Настоящая глава посвящена основным идеям и подходам асимптотической статистики интервальных данных. Приведены результаты, связанные с основополагающими в рассматриваемой области прикладной математической статистики понятиями нотны и рационального объема выборки. Рассмотрен ряд задач оценивания характеристик и параметров распределения, проверки гипотез, регрессионного, кластерного и дискриминантного анализов.

### **2.3.1. О развитии статистики интервальных данных**

Перспективная и быстро развивающаяся область статистических исследований последних лет - 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. Полученные результаты отражены, в частности, в выступлениях на проведенной в "Заводской лаборатории" дискуссии [1] и в докладах международной конференции ИНТЕРВАЛ-92 [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, а именно, заданном в техническом паспорте средства измерения, то зачастую выбросы не могут быть слишком большими. Поэтому представляются полезными иные, более общие схемы устойчивости, в частности, введенные в [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 разработке ГОСТ 11.011-83 [4] 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 2 становящиеся целыми). Ясно, что исходная длительность наработок искажена. Необходимо учесть в статистических процедурах наличие такого искажения исходных данных. Как это сделать?

Первое, что приходит в голову - модель группировки данных, согласно которой для истинного значения  $X$  проводится замена на ближайшее число из множества  $\{0,5n, n=1,2,3,\dots\}$ . Однако эту модель целесообразно подвергнуть сомнению, а также рассмотреть иные модели. Так, возможно, что  $X$  надо приводить к ближайшему сверху элементу указанного множества - если проверка качества поставленных на испытание резцов проводилась раз в полчаса. Другой вариант: если расстояния от  $X$  до двух ближайших

элементов множества  $\{0,5n, n=1,2,3,\dots\}$  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 определяется по формуле  $[a,b] - [c,d] = [a-d, b-c]$ . Для положительных  $a, b, c, d$  произведение определяется формулой  $[a,b] * [c,d] = [ac, bd]$ , 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  $[c,d]$ . Каково минимальное и максимальное значение для  $x+y$ ? Очевидно,  $a+c$  и  $b+d$  соответственно. Минимальные и максимальные значения для  $x-y, xy, x/y$  задают 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 видно из сборника трудов Международной конференции [2], к настоящему времени удалось решить, в частности, 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, есть 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 (подробнее см. ниже).

В настоящей главе развиваем 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, еще в начале 1980-х годов с помощью такой асимптотики были сформулированы правила выбора метода оценивания в ГОСТ 11.011-83 [4].

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 т.ч. 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]. 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 [12,13]. Начата разработка интервального дискриминантного анализа, в частности, рассмотрено влияние интервальности данных на показатель качества классификации [12,14]. Основные идеи и результаты рассматриваемого направления в статистике интервальных данных приведены в публикациях обзорного характера [5,6].

Как показала, в частности, международная конференция ИНТЕРВАЛ-92, 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 т.н. 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] был 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 опубликовано лишь через 5 лет в статье [10].

Следует отметить, что хотя в 1982 г. при разработке стандарта [4] были сформулированы основные идеи статистики интервальных данных, однако из-за недостатка времени они не были полностью реализованы в ГОСТ 11.011-83, и этот стандарт написан в основном в классической манере. Развитие идей статистики интервальных данных продолжается уже в течение 20 лет, и еще много чего надо сделать! Большое значение статистики интервальных данных для современной прикладной статистики обосновано в [15,16].

Ведущая научная школа в области статистики интервальных данных - это школа проф. А.П. Воцинина, активно работающая с конца 70-х годов. Полученные результаты отражены в ряде монографий (см., в частности, [17,18,19]), статей [1, 20, 21],

докладов, в частности, в трудах [2] Международной конференции ИНТЕРВАЛ-92, диссертаций [22,23]. В частности, изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности. Рассматриваемое ниже направление отличается нацеленностью на асимптотические результаты, полученные при больших объемах выборок и малых погрешностях измерений, поэтому оно и названо **асимптотической статистикой интервальных данных**.

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

### **2.3.2. Основные идеи асимптотической математической статистики интервальных данных**

Пусть существо реального явления описывается выборкой  $x_1, x_2, \dots, x_n$ . В вероятностной теории математической статистики, из которой мы исходим (см. терминологическую статью [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  $x_1, x_2, \dots, x_n$ , а величины

$$y_j = x_j + e_j, \quad j = 1, 2, \dots, n,$$

где  $e_1, e_2, \dots, e_n$  – 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  $x_1, x_2, \dots, x_n$ , в то время как статистической обработке доступны лишь искаженные значения  $y_j = x_j + e_j$ .

Введем обозначения

$$x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n), e = e_1 + e_2 + \dots + e_n.$$

Пусть статистические выводы основываются на статистике  $f: R^n \rightarrow R^1$ , 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  $f(y)$ , НО НЕ  $f(x)$ .

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

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

наблюдений  $e$ , известного статистику значения  $f(y)$  от истинного значения  $f(x)$ , т.е.

$$Nf(x) = \sup |f(y) - f(x)|,$$

где супремум берется по множеству возможных значений вектора погрешностей  $e$  (см. ниже), будем называть НОТНОЙ.

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

$$|e_i| \leq \Delta, \quad i = 1, 2, \dots, n, \quad (1)$$

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

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

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

$$e_i = \begin{cases} \Delta, & \frac{\partial f(x)}{\partial x_i} \geq 0, \\ -\Delta, & \frac{\partial f(x)}{\partial x_i} < 0, \end{cases}$$

а минимум, отличающийся от максимума только знаком, достигается при  $e_i' = -e_i$ . Следовательно, нотна с точностью до бесконечно малых более высокого порядка имеет вид

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

Это выражение назовем *асимптотической нотной*.

Условие (1) означает, что исходные данные представляются статистику в виде интервалов  $[y_i - \Delta; y_i + \Delta], i = 1, 2, \dots, n$  (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  $\Delta$ , а предельная относительная погрешность  $d$ , т.е. ограничения на погрешности вошедших в выборку результатов измерений имеют вид

$$|e_i| \leq d |x_i|, i = 1, 2, \dots, n,$$

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

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

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

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

$$\lim_{n \rightarrow \infty} N_f(x) = C\Delta. \quad (2)$$

Соотношение (2) доказывается отдельно для каждой конкретной задачи.

При использовании классических эконометрических методов в большинстве случаев используемая статистика  $f(x)$  является

асимптотически нормальной. Это означает, что существуют константы  $a$  и  $s^2$  такие, что

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

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

$$\lim_{n \rightarrow \infty} \sqrt{n}(Mf(x) - a) = 0$$

и

$$\lim_{n \rightarrow \infty} nDf(x) = s^2,$$

а потому в классической эконометрике средний квадрат ошибки статистической оценки равен

$$M(f(x) - a)^2 = (Mf(x) - a)^2 + Df(x) = \frac{s^2}{n}$$

с точностью до членов более высокого порядка.

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

$$\max_{\{e\}} M(f(y) - a)^2 = \frac{s^2}{n} + N_f^2(y) + o(\Delta^2 + \frac{1}{n}). \quad (3)$$

Из соотношения (3) можно сделать ряд важных следствий. Прежде всего отметим, что правая часть этого равенства, в отличие от правой части соответствующего классического равенства, не стремится к 0 при безграничном возрастании объема выборки. Она остается больше некоторого положительного числа, а именно, квадрата нотны. Следовательно, статистика  $f(x)$  не является состоятельной оценкой параметра  $a$ . Более того, состоятельных оценок вообще не существует.

Пусть доверительным интервалом для параметра  $a$ , соответствующим заданной доверительной вероятности  $g$ , в классической математической статистике является интервал  $(c_n(g); d_n(g))$ . В статистике интервальных данных аналогичный

доверительный интервал является более широким. Он имеет вид  $(c_n(g) - N_f(y); d_n(g) + N_f(y))$ . Таким образом, его длина увеличивается на две нотны. Следовательно, при увеличении объема выборки длина доверительного интервала не может стать меньше, чем  $2\text{СД}$  (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 - см. формулу (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", выдвинутой 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

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

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 можно "в двух словах" сформулировать так: для любого алгоритма анализа данных (алгоритма прикладной статистики) необходимо вычислить нотну и рациональный объем выборки. Или иные величины из того же понятийного ряда, возникающие в многомерном случае, при наличии нескольких выборок и при иных обобщениях описываемой здесь простейшей схемы. Затем 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  $\Delta = 0$ .

### **2.3.3. Интервальные данные в задачах оценивания характеристик распределения**

Поясним теоретические концепции статистики интервальных данных на простых примерах.

**Пример 1. Оценивание математического ожидания.** Пусть необходимо оценить математическое ожидание случайной

величины с помощью обычной оценки - среднего арифметического результатов наблюдений, т.е.

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

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

Поскольку

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

то в обозначениях предыдущего пункта

$$s^2 = D(x_1).$$

Следовательно, рациональный объем выборки равен

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

Для практического использования полученной формулы надо оценить дисперсию результатов наблюдений. Можно доказать, что, поскольку  $\Delta$  мало, это можно сделать обычным способом, например, с помощью несмещенной выборочной оценки дисперсии

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

Здесь и далее рассуждения часто идут 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  $D(x_I)$ .

Итак, выборочной оценкой рационального объема выборки является

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

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  $\Delta = \frac{S}{6}$ , то  $n_{\text{rat}} = 36$ . А 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  $g$ ) имеет другой вид:

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

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

По поводу формулы (4) была довольно жаркая дискуссия среди специалистов. Отмечалось, 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, в качестве

$u(g)$  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 среди специалистов. Выяснилось, что распределения встречающихся в практических задачах результатов измерений почти всегда отличны от нормальных [25]. А также и 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], [31, с.121] и др.).

**Пример 2. 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, что случайная величина

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

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

Из сказанного вытекает, что в статистике интервальных данных асимптотический доверительный интервал для дисперсии  $s^2$  (соответствующий доверительной вероятности  $g$ ) имеет вид

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

где

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

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

Рациональный объем выборки при оценивании дисперсии равен

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

а выборочную оценку рационального объема выборки  $n_{sample-rat}$  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  $s^2$ , то в результате вычисления моментов случайных величин в предыдущей формуле получаем, что

$$n_{rat} = \frac{s^2}{p\Delta^2},$$

где  $p$  - отношение длины окружности к диаметру,  $p = 3,141592...$   
 Например, если  $\Delta = s/6$ , то  $n_{rat} = 11$ . Это меньше, чем при  
 оценивании математического ожидания в предыдущем примере.

**Пример 3. Аддитивные статистики.** Пусть  $g : R^1 \rightarrow R^1$  -  
 некоторая непрерывная функция. Аддитивные статистики имеют  
 вид

$$f(x) = \frac{1}{n} \sum_{1 \leq i \leq n} g(x_i).$$

Тогда

$$\sum_{1 \leq i \leq n} \left| \frac{\partial f(x)}{\partial x_i} \right| = \frac{1}{n} \sum_{1 \leq i \leq n} \left| \frac{dg(x_i)}{dx_i} \right| \rightarrow M \left| \frac{dg(x_1)}{dx_1} \right|,$$

$$\sum_{1 \leq i \leq n} \left| x_i \frac{\partial f(x)}{\partial x_i} \right| = \frac{1}{n} \sum_{1 \leq i \leq n} \left| x_i \frac{dg(x_i)}{dx_i} \right| \rightarrow M \left| x_1 \frac{dg(x_1)}{dx_1} \right|$$

по вероятности при  $n \rightarrow \infty$ , если математические ожидания в правых  
 частях двух последних соотношений существуют. Применяя  
 рассмотренные выше общие соображения, получаем, что при малых  
 фиксированных  $\Delta$  и  $d$  и достаточно больших  $n$  значения  $f(y)$  могут  
 принимать любые величины из разрешенных (например,  
 записываемых заданным числом значащих цифр) в замкнутом  
 интервале

$$\left[ f(x) - \Delta M \left| \frac{dg(x_1)}{dx_1} \right|; f(x) + \Delta M \left| \frac{dg(x_1)}{dx_1} \right| \right] \quad (5)$$

при ограничениях (1) на абсолютные ошибки и в замкнутом  
 интервале

$$\left[ f(x) - dM \left| x_1 \frac{dg(x_1)}{dx_1} \right|; f(x) + dM \left| x_1 \frac{dg(x_1)}{dx_1} \right| \right] \dots (6)$$

при ограничениях на относительные погрешности результатов  
 наблюдений. Обратим внимание, что длины этих интервалов  
 независимы от объема выборки, в частности, не стремятся к 0 при  
 его росте.

К каким последствиям это приводит в задачах статистического оценивания? Поскольку для статистик аддитивного типа

$$f(x) = \frac{1}{n} \sum_{1 \leq i \leq n} g(x_i) \rightarrow Mg(x_1) \quad (7)$$

по вероятности при  $n \rightarrow \infty$ , если математическое ожидание в правой части формулы (7) существует, то аддитивную статистику  $f(x)$  естественно рассматривать как непараметрическую оценку этого математического ожидания. Термин «непараметрическая» означает, что не делается предположений о принадлежности функции распределения выборки к тому или иному параметрическому семейству распределения. Распределение статистики  $f(x)$  зависит от распределения результатов наблюдений. Однако для любого распределения результатов наблюдений с конечной дисперсией статистика  $f(x)$  является состоятельной и асимптотически нормальной оценкой для математического ожидания, указанного в правой части формулы (7).

Как известно, в рамках классической математической статистики в предположении существования ненулевой дисперсии  $Dg(x_1)$  в силу асимптотической нормальности аддитивной статистики  $f(x)$  асимптотический доверительный интервал, соответствующий доверительной вероятности  $g$ , имеет вид

$$\left[ f(x) - u \left( \frac{1+g}{2} \right) \frac{s(g(x))}{\sqrt{n}}; f(x) + u \left( \frac{1+g}{2} \right) \frac{s(g(x))}{\sqrt{n}} \right],$$

где  $s(g(x))$  – выборочное среднее квадратическое отклонение, построенное по  $g(x_1), g(x_2), \dots, g(x_n)$ , а  $u \left( \frac{1+g}{2} \right)$  – квантиль стандартного нормального распределения порядка  $\frac{1+g}{2}$ .

В рассматриваемой модели порождения интервальных данных вместо  $f(x)$  необходимо использовать  $f(y)$ , а вместо  $g(x_i)$  –

соответственно  $g(y_i)$ ,  $i=1,2,\dots,n$ . При этом доверительный интервал необходимо расширить с учетом формул (5) и (6).

В соответствии с проведенными рассуждениями для аддитивных статистик асимптотическая нотна имеет вид

$$N_f(x) = \Delta M \left| \frac{dg(x_1)}{dx_1} \right|$$

при ограничениях (1) на абсолютную погрешность и

$$N_f(x) = dM \left| x_1 \frac{dg(x_1)}{dx_1} \right|$$

при ограничениях на относительную погрешность. В первом случае нотна является обобщением понятия предельной абсолютной систематической ошибки, во втором – предельной относительной систематической ошибки. Отметим, что, как и в примерах 1 и 2, асимптотическая нотна не зависит от точки, в которой вычисляется. Таким образом, она является константой для конкретного метода статистического анализа данных.

Поскольку  $n$  велико, а  $\Delta$  и  $d$  малы, то можно пренебречь отличием выборочного среднего квадратического отклонения  $s(g(y))$ , вычисленного по выборке преобразованных значений  $g(y_1), g(y_2), \dots, g(y_n)$ , от выборочного среднего квадратического отклонения  $s(g(x))$ , построенного по выборке  $g(x_1), g(x_2), \dots, g(x_n)$ . Разность этих двух величин является бесконечно малой, они приближаются к одной и той же положительной константе.

В статистике интервальных данных выборочный доверительный интервал для  $Mg(x_1)$  имеет вид

$$\left[ f(y) - N_f(y) - u \left( \frac{1+g}{2} \right) \frac{s(g(y))}{\sqrt{n}}; f(y) + N_f(y) + u \left( \frac{1+g}{2} \right) \frac{s(g(y))}{\sqrt{n}} \right].$$

В асимптотике его длина такова:

$$2N_f(x) + 2u \left( \frac{1+d}{2} \right) \frac{s}{\sqrt{n}}, \quad (8)$$

где  $s^2$  - дисперсия  $g(x_1)$ , в то время как в классической теории математической статистики имеется только второе слагаемое. Соотношение (8) – аналог суммарной ошибки у метрологов [26]. Поскольку первое слагаемое положительно, то оценивание  $Mg(x_1)$  с помощью  $f(y)$  не является состоятельным.

Для аддитивных статистик при больших  $n$  максимум (по возможным погрешностям) среднего квадрата отклонения оценки имеет вид

$$\max_e M[f(y) - Mg(x_1)]^2 = N_f^2(x) + \frac{Dg(x_1)}{n} \quad (9)$$

с точностью до членов более высокого порядка. Исходя из принципа уравнивания погрешностей в общей схеме устойчивости [3], нецелесообразно второе слагаемое в (9) делать меньше первого за счет увеличения объема выборки  $n$ . Рациональный объем выборки, т.е. тот объем, при котором равны погрешности оценивания (или проверки гипотез), вызванные погрешностями исходных данных, и статистические погрешности, рассчитанные по обычным правилам математической статистики (при  $e_i \equiv 0$ ), для аддитивных статистик согласно (9) имеет вид

$$n_{rat} = \frac{Dg(x_1)}{N_f^2(x)}. \quad (10)$$

В качестве примера рассмотрим экспоненциально распределенные результаты наблюдений  $x_i, M(x_1) = D(x_1) = 1$ . Оцениваем математическое ожидание с помощью выборочного среднего арифметического при ограничениях на относительную погрешность. Тогда согласно формуле (10)

$$N_f(x) = d, \quad n_{rat} = \frac{1}{d^2}.$$

В частности, если относительная погрешность измерений  $d = 10\%$ , то рациональный объем выборки равен 100. Формуле (10) соответствует также рассмотренный выше пример 1.

**Пример 4. Оценивание медианы распределения с помощью выборочной медианы.** Хотя нельзя выделить главный линейный член из-за недифференцируемости функции  $f(x)$ , выражающей выборочную медиану через элементы выборки, непосредственно из определения нотны следует, что при ограничениях на абсолютные погрешности

$$N_f(x) = \Delta,$$

а при ограничениях на относительные погрешности

$$N_f(x) = dx_{med}$$

с точностью до бесконечно малых более высокого порядка, где  $x_{med}$  - теоретическая медиана. Доверительный интервал для медианы имеет вид

$$[a_1(x) - N_f(x); a_2(x) + N_f(x)],$$

где  $[a_1(x); a_2(x)]$  - доверительный интервал для медианы, вычисленный по классическим правилам непараметрической статистики [27]. Для нахождения рационального объема выборки можно использовать асимптотическую дисперсию выборочной медианы. Она, как известно (см., например, [28, с.178]), равна

$$s^2(M) = \frac{1}{4np^2(x_{med})}.$$

где  $p(x_{med})$  - плотность распределения результатов измерений в точке  $x_{med}$ . Следовательно, рациональный объем выборки имеет вид

$$n_{rat} = \frac{1}{4p^2(x_{med})\Delta^2}, \quad n_{rat} = \frac{1}{4p^2(x_{med})x_{med}^2 d^2}$$

при ограничениях на абсолютные и относительные погрешности результатов измерений соответственно. Для практического использования этих формул следует оценить плотность распределения результатов измерений в одной точке - теоретической медиане. Это можно сделать с помощью тех или иных непараметрических оценок плотности [27].

Если результаты наблюдений имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то

$$n_{rat} = \frac{p}{2\Delta^2} \approx \frac{1,57}{\Delta^2}.$$

В этом случае рациональный объем выборки в  $p/2$  раз больше, чем для оценивания математического ожидания (пример 1 выше). Однако для других распределений рассматриваемое соотношение объемов может быть иным, в частности, меньше 1. Как вытекает из статьи А.Н.Колмогорова 1931 г. [29], рассматриваемое соотношение объемов может принимать любое значение между 0 и 3.

### Пример 5. Оценивание коэффициента вариации.

Рассмотрим выборочный коэффициент вариации

$$v = f(y_1, y_2, \dots, y_n) = \frac{\left\{ \frac{1}{n-1} \sum_{1 \leq i \leq n} (y_i - \bar{y})^2 \right\}^{1/2}}{\frac{1}{n} \sum_{1 \leq i \leq n} y_i} = \frac{s(y)}{\bar{y}}.$$

Как нетрудно подсчитать,

$$\frac{\partial f}{\partial x_i} = \frac{n\bar{x}(x_i - \bar{x}) - (n-1)s^2(x)}{n(n-1)(\bar{x})^2 s(x)}.$$

В случае ограничений на относительную погрешность

$$\lim_{n \rightarrow \infty} N_f(x) = \frac{d}{(M(x_1))^2 s} M | x_1 \{ [x_1 - M(x_1)] M(x_1) - s^2 \} |.$$

На основе этого предельного соотношения и формулы для асимптотической дисперсии выборочного коэффициента вариации, приведенной в [27], могут быть найдены по описанной выше схеме доверительные границы для теоретического коэффициента вариации и рациональный объем выборки.

**Замечание.** Отметим, что формулы для рационального объема выборки получены на основе асимптотической теории, а применяются для получения конечных объемов – 36 и 100 в примерах 1-3. Как всегда при использовании асимптотических

результатов математической статистики, необходимы дополнительные исследования для изучения точности асимптотических формул при конечных объемах выборок.

#### **2.3.4. Интервальные данные в задачах оценивания параметров (на примере гамма-распределения)**

Рассмотрим классическую в прикладной математической статистике параметрическую задачу оценивания. Исходные данные – выборка  $x_1, x_2, \dots, x_n$ , состоящая из  $n$  действительных чисел. В вероятностной модели простой случайной выборки ее элементы  $x_1, x_2, \dots, x_n$  считаются набором реализаций  $n$  независимых одинаково распределенных случайных величин. Будем считать, что эти величины имеют плотность  $f(x)$ . В параметрической статистической теории предполагается, что плотность  $f(x)$  известна с точностью до конечномерного параметра, т.е.,  $f(x) = f(x, q_0)$  при некотором  $q_0 \in \Theta \subseteq R^k$ . Это, конечно, весьма сильное предположение, которое требует обоснования и проверки; однако в настоящее время параметрическая теория оценивания широко используется в различных прикладных областях.

Все результаты наблюдений определяются с некоторой точностью, в частности, записываются с помощью конечного числа значащих цифр (обычно 2 – 5). Следовательно, все реальные распределения результатов наблюдений дискретны. Обычно считают, что эти дискретные распределения достаточно хорошо приближаются непрерывными. Уточняя это утверждение, приходим к уже рассматривавшейся модели, согласно которой статистику доступны лишь величины

$$y_j = x_j + e_j, \quad j = 1, 2, \dots, n,$$

где  $x_i$  – «истинные» значения,  $e_1, e_2, \dots, e_n$  – погрешности наблюдений (включая погрешности дискретизации). В вероятностной модели принимаем, что  $n$  пар

$$(x_1, e_1), (x_2, e_2), \dots, (x_n, e_n)$$

образуют простую случайную выборку из некоторого двумерного распределения, причем  $x_1, x_2, \dots, x_n$  – выборка из распределения с плотностью  $f(x) = f(x, q_0)$ . Необходимо учитывать, что  $x_i$  и  $e_i$  – реализации зависимых случайных величин (если считать их независимыми, то распределение  $y_i$  будет непрерывным, а не дискретным). Поскольку систематическую ошибку, как правило, нельзя полностью исключить [26, с.141], то необходимо рассматривать случай  $Me_i \neq 0$ . Нет оснований априори принимать и нормальность распределения погрешностей (согласно сводкам экспериментальных данных о разнообразии форм распределения погрешностей измерений, приведенным в [26, с.148] и [27, с.71-77], в подавляющем большинстве случаев гипотеза о нормальном распределении погрешностей оказалась неприемлемой для средств измерений различных типов). Таким образом, все три распространенных представления о свойствах погрешностей не адекватны реальности. Влияние погрешностей наблюдений на свойства статистических моделей необходимо изучать на основе иных моделей, а именно, моделей интервальной статистики.

Пусть  $e$  – характеристика величины погрешности, например, средняя квадратическая ошибка  $e = \sqrt{M(e_i^2)}$ . В классической математической статистике  $e$  считается пренебрежимо малой ( $e \rightarrow 0$ ) при фиксированном объеме выборки  $n$ . Общие результаты доказываются в асимптотике  $n \rightarrow \infty$ . Таким образом, в классической математической статистике сначала делается предельный переход  $e \rightarrow 0$ , а затем предельный переход  $n \rightarrow \infty$ . В статистике интервальных данных принимаем, что объем выборки достаточно

велик ( $n \rightarrow \infty$ ), но всем измерениям соответствует одна и та же характеристика погрешности  $e \neq 0$ . Полезные для анализа реальных данных предельные теоремы получаем при  $e \rightarrow 0$ . В статистике интервальных данных сначала делается предельный переход  $n \rightarrow \infty$ , а затем предельный переход  $e \rightarrow 0$ . Итак, в обеих теориях используются одни и те же два предельных перехода:  $n \rightarrow \infty$  и  $e \rightarrow 0$ , но в разном порядке. Утверждения обеих теорий принципиально различны.

Изложение ниже идет на примере оценивания параметров гамма-распределения, хотя аналогичные результаты можно получить и для других параметрических семейств, а также для задач проверки гипотез (см. ниже) и т.д. Наша цель – продемонстрировать основные черты подхода статистики интервальных данных. Его разработка была стимулирована подготовкой ГОСТ 11.011-83 [4].

Отметим, что постановки статистики объектов нечисловой природы соответствуют подходу, принятому в общей теории устойчивости [3,27]. В соответствии с этим подходом выборке  $x = (x_1, x_2, \dots, x_n)$  ставится в соответствие множество допустимых отклонений  $G(x)$ , т.е. множество возможных значений вектора результатов наблюдений  $y = (y_1, y_2, \dots, y_n)$ . Если известно, что абсолютная погрешность результатов измерений не превосходит  $\Delta$ , то множество допустимых отклонений имеет вид

$$G(x, \Delta) = \{y : |y_i - x_i| \leq \Delta, i = 1, 2, \dots, n\}.$$

Если известно, что относительная погрешность не превосходит  $d$ , то множество допустимых отклонений имеет вид

$$G(x, d) = \{y : |\frac{y_i}{x_i} - 1| \leq d, i = 1, 2, \dots, n\}.$$

Теория устойчивости позволяет учесть «наихудшие» отклонения, т.е. приводит к выводам типа минимаксных, в то время как

конкретные модели погрешностей позволяют делать заключения о поведении статистик «в среднем».

**Оценки параметров гамма-распределения.** Как известно, случайная величина  $X$  имеет гамма-распределение, если ее плотность такова [4]:

$$f(x; a, b) = \begin{cases} \frac{1}{\Gamma(a)} x^{a-1} b^{-a} \exp\{-\frac{x}{b}\}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

где  $a$  – параметр формы,  $b$  – параметр масштаба,  $\Gamma(a)$  - гамма-функция. Отметим, что есть и иные способы параметризации семейства гамма-распределений [30].

Поскольку  $M(X) = ab$ ,  $D(X) = ab^2$ , то оценки метода имеют вид

$$\hat{a} = \frac{(\bar{x})^2}{s^2}, \quad \hat{b} = \frac{\bar{x}}{\hat{a}} = \frac{s^2}{\bar{x}},$$

где  $\bar{x}$  - выборочное среднее арифметическое, а  $s^2$  – выборочная дисперсия. Можно показать, что при больших  $n$

$$M(\hat{a} - a)^2 = \frac{2a(a+1)}{n}, \quad M(\hat{b} - b)^2 = \frac{b^2}{n} \left(2 + \frac{3}{a}\right) \quad (11)$$

с точностью до бесконечно малых более высокого порядка.

Оценка максимального правдоподобия  $a^*$  имеет вид [4]:

$$a^* = H\left(\frac{1}{n} \sum_{1 \leq i \leq n} \ln\left(\frac{\bar{x}}{x_i}\right)\right), \quad (12)$$

где  $H(\bullet)$  - функция, обратная к функции

$$Q(a) = \ln a - \frac{d\Gamma(a)}{da} / \Gamma(a).$$

При больших  $n$  с точностью до бесконечно малых более высокого порядка

$$M(a^* - a)^2 = \frac{a}{n(ay'(a) - 1)}, \quad y(a) = \frac{d\Gamma(a)}{da} / \Gamma(a).$$

Как и для оценок метода моментов, оценка максимального правдоподобия  $b^*$  параметра масштаба имеет вид

$$b^* = \bar{x}/a^*.$$

При больших  $n$  с точностью до бесконечно малых более высокого порядка

$$M(b^* - b)^2 = \frac{b^2 y'(a)}{n(a y'(a) - 1)}.$$

Используя свойства гамма-функции, можно показать [4], что при больших  $a$

$$M(a^* - a)^2 = \frac{a(2a - 1)}{n}, \quad M(b^* - b)^2 = \frac{2b^2}{n}.$$

с точностью до бесконечно малых более высокого порядка. Сравнивая с формулами (11), убеждаемся в том, что средние квадраты ошибок для оценок метода моментов больше соответствующих средних квадратов ошибок для оценок максимального правдоподобия. Таким образом, с точки зрения классической математической статистики оценки максимального правдоподобия имеют преимущество по сравнению с оценками метода моментов.

**Необходимость учета погрешностей измерений.** Положим

$$v = f(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{1 \leq i \leq n} \ln \left( \frac{\bar{x}}{x_i} \right)$$

Из свойств функции  $H(\bullet)$  следует [4, с.14], что при малых  $v$

$$a^* \sim 1/(2v). \quad (13)$$

В силу состоятельности оценки максимального правдоподобия  $a^*$  из формулы (13) следует, что  $v \rightarrow 0$  по вероятности при  $a \rightarrow \infty$ .

Согласно модели статистики интервальных данных результатами наблюдений являются не  $x_i$ , а  $y_i$ , вместо  $v$  по реальным данным рассчитывают

$$w = f(y_1, y_2, \dots, y_n) = \frac{1}{n} \sum_{1 \leq i \leq n} \ln \left( \frac{\bar{y}}{y_i} \right)$$

Имеем

$$w - v = \ln\left(\frac{\bar{y}}{\bar{x}}\right) - \frac{1}{n} \sum_{1 \leq i \leq n} \ln\left(1 + \frac{e_i}{x_i}\right) \quad (14)$$

В силу закона больших чисел при достаточно малой погрешности  $e$ , обеспечивающей возможность приближения  $\ln(1+a) \sim a$  для слагаемых в формуле (14), или, что эквивалентно, при достаточно малых предельной абсолютной погрешности  $\Delta$  в формуле (1) или достаточно малой предельной относительной погрешности  $d$  имеем при  $n \rightarrow \infty$

$$w - v \rightarrow \frac{M(e_i)}{M(x_i)} - M\left(\frac{e_i}{x_i}\right) = c$$

по вероятности (в предположении, что все погрешности одинаково распределены). Таким образом, наличие погрешностей вносит сдвиг, вообще говоря, не исчезающий при росте объема выборки. Следовательно, если  $c \neq 0$ , то оценка максимального правдоподобия не является состоятельной. Имеем

$$a^*(y) - a^* \approx -\frac{c}{2v^2},$$

где величина  $a^*(y)$  определена по формуле (12) с заменой  $x_i$  на  $y_i$ ,  $i=1,2,\dots,n$ . Из формулы (13) следует [4], что

$$a^*(y) - a \approx -2(a^*)^2 c, \quad (15)$$

т.е. влияние погрешностей измерений увеличивается по мере роста  $a$ .

Из формул для  $v$  и  $w$  следует, что с точностью до бесконечно малых более высокого порядка

$$w - v \approx \sum_{1 \leq i \leq n} \frac{\partial f}{\partial x_i} e_i = \frac{1}{n} \sum_{1 \leq i \leq n} \left( \frac{1}{\bar{x}} - \frac{1}{x_i} \right) e_i. \quad (16)$$

С целью нахождения асимптотического распределения  $w$  выделим, используя формулу (16) и формулу для  $v$ , главные члены в соответствующих слагаемых

$$w = \ln M(x_1) + \frac{1}{n} \sum_{1 \leq i \leq n} \left\{ \frac{x_i - M(x_1)}{M(x_1)} - \ln x_i + \left( \frac{1}{M(x_1)} - \frac{1}{x_i} \right) e_i \right\} + O_p\left(\frac{1}{n}\right). \quad (17)$$

Таким образом, величина  $w$  представлена в виде суммы независимых одинаково распределенных случайных величин (с точностью до зависящего от случая остаточного члена порядка  $1/n$ ). В каждом слагаемом выделяются две части – одна, соответствующая  $M_b$  и вторая, в которую входят  $e_i$ . На основе представления (17) можно показать, что при  $n \rightarrow \infty, e \rightarrow 0$  распределения случайных величин  $v$  и  $w$  асимптотически нормальны, причем

$$M(w) \approx M(v) + c, \quad D(w) \approx D(v).$$

Из асимптотического совпадения дисперсий  $v$  и  $w$ , вида параметров асимптотического распределения (при  $a \rightarrow \infty$ ) оценки максимального правдоподобия  $a^*$  и формулы (15) вытекает одно из основных соотношений статистики интервальных данных

$$M(a^*(y) - a)^2 \approx 4a^4 c^2 + \frac{a(2a-1)}{n}. \quad (18)$$

Соотношение (18) уточняет утверждение о несостоятельности  $a^*$ . Из него следует также, что не имеет смысла безгранично увеличивать объем выборки  $n$  с целью повышения точности оценивания параметра  $a$ , поскольку при этом уменьшается только второе слагаемое в (18), а первое остается постоянным.

В соответствии с общим подходом статистики интервальных данных в стандарте [4] предлагается определять рациональный объем выборки  $n_{rat}$  определять из условия «уравнивания погрешностей» (предложено в монографии [3]) различных видов в формуле (18), т.е. из условия

$$4a^4 c^2 = \frac{a(2a-1)}{n_{rat}}.$$

Упрощая это уравнение в предположении  $a \rightarrow \infty$ , получаем, что

$$n_{rat} = \frac{1}{2a^2 c^2}.$$

Согласно сказанному выше, целесообразно использовать лишь выборки с объемами  $n \leq n_{rat}$ . Превышение рационального объема выборки  $n_{rat}$  не дает существенного повышения точности оценивания.

**Применение методов теории устойчивости.** Найдем асимптотическую нотну. Как следует из вида главного линейного члена в формуле (17), решение оптимизационной задачи

$$w - v \rightarrow \max, \quad |e_i| \leq \Delta,$$

соответствующей ограничениям на абсолютные погрешности, имеет вид

$$e_i = \begin{cases} \Delta, & \frac{1}{\bar{x}} - \frac{1}{x_i} \geq 0, \\ -\Delta, & \frac{1}{\bar{x}} - \frac{1}{x_i} < 0 \end{cases}.$$

Однако при этом пары  $(x_i, e_i)$  не образуют простую случайную выборку, т.к. в выражения для  $e_i$  входит  $\bar{x}$ . Однако при  $n \rightarrow \infty$  можно заменить  $\bar{x}$  на  $M(x_1)$ . Тогда получаем, что

$$w - v \approx A\Delta$$

при  $a > 1$ , где

$$A = M \left| \frac{1}{M(x_1)} - \frac{1}{x_1} \right| = \int_0^{\infty} \left| \frac{1}{ab} - \frac{1}{x} \right| f(x; a, b) dx.$$

Таким образом, с точностью до бесконечно малых более высокого порядка нотна имеет вид

$$N_{a^*}(y) = 2(a^*)^2 c, \quad c = A\Delta.$$

Применим полученные результаты к построению доверительных интервалов. В постановке классической математической статистики (т.е. при  $e = 0$ ) доверительный интервал для параметра формы  $a$ , соответствующий доверительной вероятности  $g$ , имеет вид [4]

$$\left[ a^* - u \left( \frac{1+g}{2} \right) \mathfrak{S}^*(a^*); a^* + u \left( \frac{1+g}{2} \right) \mathfrak{S}^*(a^*) \right],$$

где  $u \left( \frac{1+g}{2} \right)$  - квантиль порядка  $\frac{1+g}{2}$  стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1,

$$[\mathfrak{S}^*(a^*)]^2 = \frac{a^*}{n(a^* y'(a^*) - 1)}, \quad y(a) = \frac{d\Gamma(a)}{da} / \Gamma(a).$$

В постановке статистики интервальных данных (т.е. при  $e \neq 0$ ) следует рассматривать доверительный интервал

$$\left[ a^* - 2(a^*)^2 |c| - u \left( \frac{1+g}{2} \right) \mathfrak{S}^*(a^*); a^* + 2(a^*)^2 |c| + u \left( \frac{1+g}{2} \right) \mathfrak{S}^*(a^*) \right],$$

где

$$c = \frac{M(e_i)}{M(x_i)} - M \left( \frac{e_i}{x_i} \right)$$

в вероятностной постановке (пары  $(x_i, e_i)$  образуют простую случайную выборку) и  $c = \Delta d$  в оптимизационной постановке. Как в вероятностной, так и в оптимизационной постановках длина доверительного интервала не стремится к 0 при  $n \rightarrow \infty$ .

Если ограничения наложены на предельную относительную погрешность, задана величина  $d$ , то значение  $c$  можно найти с помощью следующих правил приближенных вычислений [32, с.142].

- (I) Относительная погрешность суммы заключена между наибольшей и наименьшей из относительных погрешностей слагаемых.
- (II) Относительная погрешность произведения и частного равна сумме относительных погрешностей сомножителей или, соответственно, делимого и делителя.

Можно показать, что в рамках статистики интервальных данных с ограничениями на относительную погрешность правила (I) и (II) являются строгими утверждениями при  $d \rightarrow 0$ .

Обозначим относительную погрешность некоторой величины  $t$  через  $ОП(t)$ , абсолютную погрешность – через  $АП(t)$ .

Из правила (I) следует, что  $ОП(\bar{x}) = d$ , а из правила (II) – что

$$ОП\left(\frac{\bar{x}}{x_i}\right) = 2d.$$

Поскольку рассмотрения ведутся при  $a \rightarrow \infty$ , то в силу неравенства Чебышева

$$\frac{\bar{x}}{x_i} \rightarrow 1 \quad (19)$$

по вероятности при  $a \rightarrow \infty$ , поскольку и числитель, и знаменатель в (19) с близкой к 1 вероятностью лежат в промежутке  $[ab - db\sqrt{a}; ab + db\sqrt{a}]$ , где константа  $d$  может быть определена с помощью упомянутого неравенства Чебышева.

Поскольку при справедливости (19) с точностью до бесконечно малых более высокого порядка

$$\ln\left(\frac{\bar{x}}{x_i}\right) \approx \frac{\bar{x}}{x_i} - 1,$$

то с помощью трех последних соотношений имеем

$$ОП\left(\frac{\bar{x}}{x_i}\right) = АП\left(\frac{\bar{x}}{x_i}\right) = АП\left(\ln\left(\frac{\bar{x}}{x_i}\right)\right) = 2d. \quad (20)$$

Применим еще одно правило приближенных вычислений [32, с.142].

(III) Предельная абсолютная погрешность суммы равна сумме предельных абсолютных погрешностей слагаемых.

Из (20) и правила (III) следует, что

$$АП(v) = 2d. \quad (21)$$

Из (15) и (21) вытекает [4, с.44, ф-ла (18)], что

$$АП(a^*) = 4a^2d,$$

откуда в соответствии с ранее полученной формулой для рационального объема выборки с заменой  $c = 2d$  получаем, что

$$n_{rat} = \frac{1}{8a^2d^2}.$$

В частности, при  $a = 5,00$ ,  $d = 0,01$  получаем  $n_{rat} = 50$ , т.е. в ситуации, в которой были получены данные о наработке резцов до предельного состояния [4, с.29], проводить более 50 наблюдений нерационально.

В соответствии с ранее проведенными рассмотрениями асимптотический доверительный интервал для  $a$ , соответствующий доверительной вероятности  $g = 0,95$ , имеет вид

$$\left[ a^* - 4(a^*)^2 d - 1,96 \sqrt{\frac{a^*(2a^* - 1)}{n}}; \quad a^* + 4(a^*)^2 d + 1,96 \sqrt{\frac{a^*(2a^* - 1)}{n}} \right].$$

В частности, при  $a^* = 5,00$ ,  $d = 0,01$ ,  $n = 50$  имеем асимптотический доверительный интервал  $[2,12; 7,86]$  вместо  $[3,14; 6,86]$  при  $d = 0$ .

При больших  $a$  в силу соображений, приведенных при выводе формулы (19), можно связать между собой относительную и абсолютную погрешности результатов наблюдений  $x_i$  :

$$d = \frac{\Delta}{M(x_i)} = \frac{\Delta}{ab}. \quad (21)$$

Следовательно, при больших  $a$  имеем

$$c = 2d = A\Delta, \quad A = \frac{2d}{\Delta} = \frac{2}{ab}.$$

Таким образом, проведенные рассуждения дали возможность вычислить асимптотику интеграла, задающего величину  $A$ .

**Сравнение методов оценивания.** Изучим влияние погрешностей измерений (с ограничениями на абсолютную погрешность) на оценку  $\hat{a}$  метода моментов. Имеем

$$АП(\bar{x}) = \Delta, \quad АП((\bar{x})^2) \approx 2\bar{x}\Delta \approx 2ab\Delta.$$

Погрешность  $s^2$  зависит от способа вычисления  $s^2$ . Если используется формула

$$s^2 = \frac{1}{n-1} \sum_{1 \leq i \leq n} (x_i - \bar{x})^2, \quad (22)$$

то необходимо использовать соотношения

$$АП(x_i - \bar{x})^2 = 2\Delta, \quad АП[(x_i - \bar{x})^2] \approx 2|x_i - \bar{x}| \Delta.$$

По сравнению с анализом влияния погрешностей на оценку  $a^*$  здесь возникает новый момент – необходимость учета погрешностей в случайной составляющей отклонения оценки  $\hat{a}$  от оцениваемого параметра, в то время как при рассмотрении оценки максимального правдоподобия погрешности давали лишь смещение. Примем в соответствии с неравенством Чебышева

$$|x_i - \bar{x}| \sim \sqrt{D(x_1)}, \quad (23)$$

тогда

$$АП[(x_i - \bar{x})^2] \sim 2b\sqrt{a}\Delta, \quad АП(s^2) \sim 2b\sqrt{a}\Delta.$$

Если вычислять  $s^2$  по формуле

$$s^2 = \frac{1}{n-1} \sum_{1 \leq i \leq n} x_i^2 - \frac{n}{n-1} (\bar{x})^2, \quad (24)$$

то аналогичные вычисления дают, что

$$АП(s^2) \sim 4ab\Delta,$$

т.е. погрешность при больших  $a$  существенно больше. Хотя правые части формул (22) и (24) тождественно равны, но погрешности вычислений по этим формулам весьма отличаются. Связано это с тем, что в формуле (24) последняя операция – нахождение разности двух больших чисел, примерно равных по величине (для выборки из гамма-распределения при большом значении параметра формы).

Из полученных результатов следует, что

$$АП(\hat{a}) = АП\left(\frac{(\bar{x})^2}{s^2}\right) \sim \frac{2\Delta}{b}(1 + \sqrt{a}).$$

При выводе этой формулы использована линеаризация влияния погрешностей (выделение главного линейного члена). Используя связь (21) между абсолютной и относительной погрешностями, можно записать

$$АП(\hat{a}) \sim 2a(1 + \sqrt{a})d.$$

Эта формула отличается от приведенной в [4, с.44, ф-ла (19)]

$$АП(\hat{a}) \sim 2a(1 + 3\sqrt{a})d,$$

поскольку в [4] вместо (23) использовалась оценка

$$|x_i - \bar{x}| < 3\sqrt{D(x_i)}.$$

Используя соотношение (23), мы характеризуем влияние погрешностей «в среднем».

Доверительный интервал, соответствующий доверительной вероятности 0,95, имеет вид

$$[\hat{a} - 2\hat{a}(1 + \sqrt{\hat{a}})d - 1,96\sqrt{\frac{2\hat{a}(\hat{a} + 1)}{n}}; \hat{a} + 2\hat{a}(1 + \sqrt{\hat{a}})d + 1,96\sqrt{\frac{2\hat{a}(\hat{a} + 1)}{n}}].$$

Если  $\hat{a} = 5,00$ ,  $d = 0,01$ ,  $n = 50$ , то получаем доверительный интервал  $[2,54; 7,46]$  вместо  $[2,86; 7,14]$  при  $d = 0$ . Хотя при  $d = 0$  доверительный интервал для  $a$  при использовании оценки метода моментов  $\hat{a}$  шире, чем при использовании оценки максимального правдоподобия  $a^*$ , при  $d = 0,01$  результат сравнения длин интервалов противоположен.

Необходимо выбрать способ сравнения двух методов оценивания параметра  $a$ , поскольку в длины доверительных интервалов входят две составляющие – зависящая от доверительной вероятности и не зависящая от нее. Выберем  $d = 0,68$ , т.е.

$$u\left(\frac{1+g}{2}\right) = 1,00. \text{ Тогда оценке максимального правдоподобия } a^*$$

соответствует полудлина доверительного интервала

$$n(a^*) = 4a^2d + \sqrt{\frac{a(2a-1)}{n}}, \quad (25)$$

а оценке  $\hat{a}$  метода моментов соответствует полудлина доверительного интервала

$$n(\hat{a}) = 2a(1 + \sqrt{a})d + \sqrt{\frac{2a(a+1)}{n}}. \quad (26)$$

Ясно, что больших  $a$  или больших  $n$  справедливо неравенство  $n(a^*) > n(\hat{a})$ , т.е. метод моментов лучше метода максимального правдоподобия, вопреки классическим результатам Р.Фишера при  $d = 0$  [33,с.99].

Из (25) и (26) элементарными преобразованиями получаем следующее правило принятия решений. Если

$$d\sqrt{n} \geq \frac{\sqrt{2a(a+1)} - \sqrt{a(2a-1)}}{4a^2 - 2a(1+\sqrt{a})} = B(a),$$

то  $n(a^*) \geq n(\hat{a})$  и следует использовать  $\hat{a}$ ; а если  $d\sqrt{n} < B(a)$ , то  $n(a^*) < n(\hat{a})$  и надо применять  $a^*$ . Для выбора метода оценивания при обработке реальных данных целесообразно использовать  $B(\hat{a})$  (см. раздел 5 в ГОСТ 11.011-83 [4, с.10-11]).

Пример анализа реальных данных опубликован в [4].

На основе рассмотрения проблем оценивания параметров гамма-распределения можно сделать некоторые общие выводы. Если в классической теории математической статистики:

а) существуют состоятельные оценки  $a_n$  параметра  $a$ ,

$$\lim_{n \rightarrow \infty} M(a_n - a)^2 = 0;$$

б) для повышения точности оценивания объем выборки целесообразно безгранично увеличивать;

в) оценки максимального правдоподобия лучше оценок метода моментов,

то в статистике интервальных данных, учитывающей погрешности измерений, соответственно:

а) не существует состоятельных оценок: для любой оценки  $a_n$  существует константа  $c$  такая, что

$$\lim_{n \rightarrow \infty} M(a_n - a)^2 \geq c > 0;$$

б) не имеет смысла рассматривать объемы выборок, большие «рационального объема выборки»  $n_{rat}$ ;

в) оценки метода моментов в обширной области параметров  $(a, n, d)$  лучше оценок максимального правдоподобия, в частности, при  $a \rightarrow \infty$  и при  $n \rightarrow \infty$ .

Ясно, что приведенные выше результаты справедливы не только для рассмотренной задачи оценивания параметров гамма-распределения, но и для многих других постановок прикладной математической статистики.

**Метрологические, методические, статистические и вычислительные погрешности.** Целесообразно выделить ряд видов погрешностей статистических данных. Погрешности, вызванные неточностью измерения исходных данных, называем метрологическими. Их максимальное значение можно оценить с помощью нотны. Впрочем, выше на примере оценивания параметров гамма-распределения показано, что переход от максимального отклонения к реально имеющемуся в вероятностно-статистической модели не меняет выводы (с точностью до умножения предельных значений погрешностей  $\Delta$  или  $d$  на константы). Как правило, метрологические погрешности не убывают с ростом объема выборки.

Методические погрешности вызваны неадекватностью вероятностно-статистической модели, отклонением реальности от ее предпосылок. Неадекватность обычно не исчезает при росте объема выборки. Методические погрешности целесообразно изучать с помощью «общей схемы устойчивости» [3,27], обобщающей популярную в теории робастных статистических процедур модель засорения большими выбросами. В настоящей главе методические погрешности не рассматриваются.

Статистическая погрешность – это та погрешность, которая традиционно рассматривается в математической статистике. Ее характеристики – дисперсия оценки, дополнение до 1 мощности критерия при фиксированной альтернативе и т.д. Как правило, статистическая погрешность стремится к 0 при росте объема выборки.

Вычислительная погрешность определяется алгоритмами расчета, в частности, правилами округления. На уровне чистой математики справедливо тождество правых частей формул (22) и (24), задающих выборочную дисперсию  $s^2$ , а на уровне вычислительной математики формула (22) дает при определенных условиях существенно больше верных значащих цифр, чем вторая [34, с.51-52].

Выше на примере задачи оценивания параметров гамма-распределения рассмотрено совместное действие метрологических и вычислительных погрешностей, причем погрешности вычислений оценивались по классическим правилам для ручного счета [32]. Оказалось, что при таком подходе оценки метода моментов имеют преимущество перед оценками максимального правдоподобия в обширной области изменения параметров. Однако, если учитывать только метрологические погрешности, как это делалось выше в примерах 1-5, то с помощью аналогичных выкладок можно показать, что оценки этих двух типов имеют (при достаточно больших  $n$ ) одинаковую погрешность.

Вычислительную погрешность здесь подробно не рассматриваем. Ряд интересных результатов о ее роли в статистике получили Н.Н.Ляшенко и М.С.Никулин [35].

### **2.3.5. Сравнение методов оценивания параметров**

В теории оценивания параметров классической математической статистики установлено, что метод максимального правдоподобия, как правило, лучше (в смысле асимптотической дисперсии асимптотического среднего квадрата ошибки), чем метод моментов. Однако в интервальной статистике это, вообще говоря, не так, что продемонстрировано выше на примере оценивания параметров гамма-распределения. Сравним эти два метода

оценивания в случае интервальных данных в общей постановке. Поскольку метод максимального правдоподобия – частный случай метода минимального контраста, начнем с разбора этого несколько более общего метода.

**Оценки минимального контраста.** Пусть  $X$  – пространство, в котором лежат независимые одинаково распределенные случайные элементы  $x_1, x_2, \dots, x_n, \dots$ . Будем оценивать элемент пространства параметров  $\Theta$  с помощью функции контраста  $f : X \times \Theta \rightarrow R^1$ . Оценкой минимального контраста называется

$$q_n = \text{Arg min} \left\{ \sum_{1 \leq i \leq n} f(x_i, q), q \in \Theta \right\}.$$

Если множество  $q_n$  состоит из более чем одного элемента, то оценкой минимального контраста называют также любой элемент  $q_n$ .

Оценками минимального контраста являются, в частности, многие робастные статистики [3,36]. Эти оценки широко используются в статистике объектов нечисловой природы [3,27], поскольку при  $X = \Theta$  переходят в переходят в эмпирические средние, а если  $X = \Theta$  – пространство бинарных отношений – в медиану Кемени.

Пусть в  $X$  имеется мера  $m$  (заданная на той же  $\sigma$ -алгебре, что участвует в определении случайных элементов  $x_i$ ), и  $p(x; q)$  – плотность распределения  $x_i$  по мере  $m$ . Если

$$f(x; q) = -\ln p(x; q),$$

то оценка минимального контраста переходит в оценку максимального правдоподобия.

Асимптотическое поведение оценок минимального контраста в случае пространств  $X$  и  $\Theta$  общего вида хорошо изучено [37], в частности, известны условия состоятельности оценок. Здесь

ограничимся случаем  $X = R^l$ , но при этом введя погрешности измерений  $e_i$ . Примем также, что  $\Theta = (q_{\min}, q_{\max}) \subseteq R^1$ .

В рассматриваемой математической модели предполагается, что статистику известны лишь искаженные значения  $y_i = x_i + e_i$ ,  $i = 1, 2, \dots, n$ . Поэтому вместо  $q_n$  он вычисляет

$$q_n^* = \text{Arg min} \left\{ \sum_{1 \leq i \leq n} f(y_i, q), q \in \Theta \right\}.$$

Будем изучать величину  $q_n^* - q_n$  в предположении, что погрешности измерений  $e_i$  малы. Цель этого изучения – продемонстрировать идеи статистики интервальных данных при достаточно простых предположениях. Поэтому естественно следовать условиям и ходу рассуждений, которые обычно принимаются при изучении оценок максимального правдоподобия [38, п.33.3].

Пусть  $q_0$  - истинное значение параметра, функция  $f(x; q)$  трижды дифференцируема по  $q$ , причем

$$\left| \frac{\partial^3 f(x; q)}{\partial q^3} \right| < H(x)$$

при всех  $x, q$ . Тогда

$$\frac{\partial f(x; q)}{\partial q} = \frac{\partial f(x; q_0)}{\partial q} + \frac{\partial^2 f(x; q_0)}{\partial q^2} (q - q_0) + \frac{1}{2} a(x) H(x) (q - q_0)^2, \quad (27)$$

где  $|a(x)| < 1$ .

Используя обозначения векторов  $x = (x_1, x_2, \dots, x_n)$ ,  $y = (y_1, y_2, \dots, y_n)$ , введем суммы

$$B_0(x) = \frac{1}{n} \sum_{1 \leq i \leq n} \frac{\partial f(x_i; q_0)}{\partial q}, \quad B_1(x) = \frac{1}{n} \sum_{1 \leq i \leq n} \frac{\partial^2 f(x_i; q_0)}{\partial q^2}, \quad R(x) = \frac{1}{n} \sum_{1 \leq i \leq n} H(x_i).$$

Аналогичным образом введем функции  $B_0(y)$ ,  $B_1(y)$ ,  $R(y)$ , в которых вместо  $x_i$  стоят  $y_i$ ,  $i = 1, 2, \dots, n$ .

Поскольку в соответствии с теоремой Ферма оценка минимального контраста  $q_n$  удовлетворяет уравнению

$$\sum_{1 \leq i \leq n} \frac{\partial f(x_i; q_n)}{\partial q} = 0, \quad (28)$$

то, подставляя в (27)  $x_i$  вместо  $x$  и суммируя по  $i = 1, 2, \dots, n$ , получаем, что

$$0 = B_0(x) + B_1(x)(q_n - q_0) + \frac{bR(x)}{2}(q_n - q_0)^2, \quad |b| < 1, \quad (29)$$

откуда

$$q_n - q_0 = \frac{-B_0(x)}{B_1(x) + \frac{bR(x)}{2}(q_n - q_0)}. \quad (30)$$

Решения уравнения (28) будем также называть оценками минимального контраста. Хотя уравнение (28) – лишь необходимое условие минимума, такое словоупотребление не будет вызывать трудностей.

**Теорема 1.** Пусть для любого  $x$  выполнено соотношение (27). Пусть для случайной величины  $x_1$  с распределением, соответствующим значению параметра  $q = q_0$ , существуют математические ожидания

$$M \frac{\partial f(x_1, q_0)}{\partial q_0} = 0, \quad M \frac{\partial^2 f(x_1, q_0)}{\partial q_0^2} = A \neq 0, \quad MH(x_1) = M < +\infty. \quad (31)$$

Тогда существуют оценки минимального контраста  $q_n$  такие, что  $q_n \rightarrow q_0$  при  $n \rightarrow \infty$  (в смысле сходимости по вероятности).

*Доказательство.* Возьмем  $e > 0$  и  $d > 0$ . В силу закона больших чисел (теорема Хинчина) существует  $n(e, d)$  такое, что для любого  $n > n(e, d)$  справедливы неравенства

$$P\{|B_0| \geq d^2\} < e/3, \quad P\{|B_1| < |A|/2\} < e/3, \quad P\{R(x) > 2M\} < e/3.$$

Тогда с вероятностью не менее  $1 - e$  одновременно выполняются соотношения

$$|B_0| \leq d^2, \quad |B_1| \geq |A|/2, \quad R(x) \leq 2M. \quad (32)$$

При  $q \in [q_0 - d; q_0 + d]$  рассмотрим многочлен второй степени

$$y(q) = B_0(x) + B_1(x)(q - q_0) + \frac{bR(x)}{2}(q - q_0)^2$$

(см. формулу (29)). С вероятностью не менее  $1 - \epsilon$  выполнены соотношения

$$\left| B_0 + \frac{bR(x)}{2}(q - q_0)^2 \right| \leq |B_0| + \frac{R(x)d^2}{2} \leq d^2(M + 1), \quad |B_1 d| \geq \frac{|A|d}{2}.$$

Если  $0 < 2(M + 1)d < |A|$ , то знак  $y(q)$  в точках  $q_1 = q_0 - d$  и  $q_2 = q_0 + d$  определяется знаком линейного члена  $B_1(q_i - q_0)$ ,  $i = 1, 2$ , следовательно, знаки  $y(q_1)$  и  $y(q_2)$  различны, а потому существует  $q_n \in [q_0 - d; q_0 + d]$  такое, что  $y(q_n) = 0$ , что и требовалось доказать.

**Теорема 2.** Пусть выполнены условия теоремы 2 и, кроме того, для случайной величины  $x_1$ , распределение которой соответствует значению параметра  $q = q_0$ , существует математическое ожидание

$$M\left(\frac{\partial f(x_1; q_0)}{\partial q_0}\right) = s^2.$$

Тогда оценка минимального контраста имеет асимптотически нормальное распределение:

$$\lim_{n \rightarrow \infty} P\left\{\sqrt{n} \frac{|A|}{s} (q_n - q_0) < x\right\} = \Phi(x) \quad (33)$$

для любого  $x$ , где  $\Phi(x)$  – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

*Доказательство.* Из центральной предельной теоремы вытекает, что числитель в правой части формулы (30) асимптотически нормален с математическим ожиданием 0 и дисперсией  $s^2$ . Первое слагаемое в знаменателе формулы (30) в силу условий (31) и закона больших чисел сходится по вероятности к  $A \neq 0$ , а второе слагаемое по тем же основаниям и с учетом теоремы 1 – к 0. Итак, знаменатель сходится по вероятности к  $A \neq 0$ . Доказательство теоремы 2 завершает ссылка на теорему о наследовании сходимости [3, параграф 2.4].

**Нотна оценки минимального контраста.** Аналогично (30)

нетрудно получить, что

$$q_n^* - q_0 = \frac{-B_0(y)}{B_1(y) + \frac{b(y)R(y)}{2}(q_n^* - q_0)}, \quad |b(y)| < 1. \quad (34)$$

Следовательно,  $q_n^* - q_n$  есть разность правых частей формул (30) и (34). Найдем максимально возможное значение (т.е. нотну) величины  $|q_n^* - q_n|$  при ограничениях (1) на абсолютные погрешности результатов измерений.

Покажем, что при  $\Delta \rightarrow 0$  для некоторого  $C > 0$  нотна имеет вид

$$N_{q_n}(x) = \sup_{\{e\}} |q_n^* - q_n| = C\Delta(1 + o(1)). \quad (35)$$

Поскольку  $q_n^* - q_n = (q_n^* - q_0) + (q_0 - q_n)$ , то из (33) и (35) следует, что

$$\sup_{\{e\}} M(q_n^* - q_n)^2 = \left( C^2 \Delta^2 + \frac{S^2}{A^2 n} \right) (1 + o(1)). \quad (36)$$

Можно сказать, что наличие погрешностей  $e_i$  приводит к появлению систематической ошибки (смещения) у оценки метода максимального правдоподобия, и нотна является максимально возможным значением этой систематической ошибки.

В правой части (36) первое слагаемое – квадрат асимптотической нотны, второе соответствует статистической ошибке. Приравнивая их, получаем рациональный объем выборки

$$n_{rat} = \left( \frac{S}{CA\Delta} \right)^2.$$

Остается доказать соотношение (35) и вычислить  $C$ . Укажем сначала условия, при которых  $q_n^* \rightarrow q_0$  (по вероятности) при  $n \rightarrow \infty$  одновременно с  $\Delta \rightarrow 0$ .

**Теорема 3.** Пусть существуют константа  $\Delta_0$  и функции  $g_1(x)$ ,  $g_2(x)$ ,  $g_3(x)$  такие, что при  $0 \leq \Delta \leq \Delta_0$  и  $-1 \leq g \leq 1$  выполнены неравенства (ср. формулу (27))

$$\left| \frac{\partial f(x; q_0)}{\partial q} - \frac{\partial f(x + g\Delta; q_0)}{\partial q} \right| \leq g_1(x)\Delta,$$

$$\left| \frac{\partial^2 f(x; q_0)}{\partial q^2} - \frac{\partial^2 f(x + g\Delta; q_0)}{\partial q^2} \right| \leq g_2(x)\Delta, \quad \dots (37)$$

$$|H(x) - H(x + g\Delta)| \leq g_3(x)\Delta$$

при всех  $x$ . Пусть для случайной величины  $x_1$ , распределение которой соответствует  $q = q_0$ , существуют  $m_1 = Mg_1(x_1)$ ,  $m_2 = Mg_2(x_1)$  и  $m_3 = Mg_3(x_1)$ . Пусть выполнены условия теоремы 1. Тогда  $q_n^* \rightarrow q_0$  (по вероятности) при  $\Delta \rightarrow 0$ ,  $n \rightarrow \infty$ .

*Доказательство* проведем по схеме доказательства теоремы

1. Из неравенств (37) вытекает, что

$$\begin{aligned} |B_0(y) - B_0(x)| &\leq \Delta \left( \frac{1}{n} \sum_{1 \leq i \leq n} g_1(x_i) \right) \\ |B_1(y) - B_1(x)| &\leq \Delta \left( \frac{1}{n} \sum_{1 \leq i \leq n} g_2(x_i) \right) \\ |R(y) - R(x)| &\leq \Delta \left( \frac{1}{n} \sum_{1 \leq i \leq n} g_3(x_i) \right) \end{aligned} \quad (38)$$

Возьмем  $\epsilon > 0$  и  $d > 0$ . В силу закона больших чисел (теорема Хинчина) существует  $n(\epsilon, d)$  такое, что для любого  $n > n(\epsilon, d)$  справедливы неравенства

$$\begin{aligned} P \left\{ |B_0| \geq \frac{d^2}{2} \right\} &< \frac{\epsilon}{6}, \quad P \left\{ |B_1| < \frac{3|A|}{4} \right\} < \frac{\epsilon}{6}, \quad P \left\{ R(x) > \frac{3M}{2} \right\} < \frac{\epsilon}{6}, \\ P \left\{ \frac{1}{n} \sum_{1 \leq i \leq n} g_j(x_i) > 2m_j \right\} &< \frac{\epsilon}{6}, \quad j = 1, 2, 3. \end{aligned}$$

Тогда с вероятностью не менее  $1 - \epsilon$  одновременно выполняются соотношения

$$|B_0| < \frac{1}{2}d^2, \quad |B_1| \geq \frac{3|A|}{4}, \quad R(x) \leq \frac{3M}{2}, \quad \frac{1}{n} \sum_{1 \leq i \leq n} g_j(x_i) \leq 2m_j, \quad j = 1, 2, 3.$$

В силу (38) при этом

$$|B_0(y)| < \frac{1}{2}d^2 + 2\Delta m_1, \quad |B_1(y)| \geq \frac{3|A|}{4} - 2\Delta m_2, \quad R(y) \leq \frac{3M}{2} + 2\Delta m_3.$$

Пусть

$$0 \leq \Delta \leq \min \left\{ \frac{1}{4} \frac{d^2}{m_1}; \frac{1}{8} \frac{|A|}{m_2}; \frac{1}{4} \frac{M}{m_3} \right\}.$$

Тогда с вероятностью не менее  $1 - \epsilon$  одновременно выполняются соотношения (ср. (32))

$$|B_0(y)| \leq d^2, \quad |B_1(y)| \geq |A|/2, \quad R(y) \leq 2M.$$

Завершается доказательство дословным повторением такового в теореме 1, с единственным отличием – заменой в обозначениях  $x$  на  $y$ .

**Теорема 4.** Пусть выполнены условия теоремы 3 и, кроме того, существуют математические ожидания (при  $q = q_0$ )

$$M \left| \frac{\partial^2 f(x_1, q_0)}{\partial x \partial q} \right|, \quad M \left| \frac{\partial^3 f(x_1, q_0)}{\partial x \partial q^2} \right|. \quad (39)$$

Тогда выполнено соотношение (35) с

$$C = \frac{1}{|A|} M \left| \frac{\partial^2 f(x_1, q_0)}{\partial x \partial q} \right|. \quad (40)$$

*Доказательство.* Воспользуемся следующим элементарным соотношением. Пусть  $a$  и  $b$  – бесконечно малые по сравнению с  $Z$  и  $B$  соответственно. Тогда с точностью до бесконечно малых более высокого порядка

$$\frac{Z+a}{B+b} - \frac{Z}{B} = \frac{aB-bZ}{B^2}.$$

Чтобы применить это соотношение к анализу  $q_n^* - q_n$  в соответствии с (30), (34) и теоремой 2, положим

$$Z = B_0(x), \quad a = B_0(y) - B_0(x), \quad B = B_1(x), \quad b = (B_1(y) - B_1(x)) + \frac{b(y)R(y)}{2}(q_n^* - q_0).$$

В силу условий теоремы 4 при малых  $\epsilon_i$  с точностью до членов более высокого порядка

$$B_0(y) - B_0(x) = \frac{1}{n} \sum_{1 \leq i \leq n} \frac{\partial^2 f(x_i; q_0)}{\partial x_i \partial q_0} \epsilon_i, \quad B_1(y) - B_1(x) = \frac{1}{n} \sum_{1 \leq i \leq n} \frac{\partial^3 f(x_i; q_0)}{\partial x_i \partial q_0^2} \epsilon_i.$$

При  $\Delta \rightarrow 0$  эти величины бесконечно малы, а потому с учетом сходимости  $B_1(x)$  к  $A$  и теоремы 3

$$q_n^* - q_n = \frac{1}{A^2} \{(B_0(y) - B_0(x))A - (B_1(y) - B_1(x))B_0(x)\} = \frac{1}{A^2 n} \sum_{1 \leq i \leq n} g_i e_i$$

с точностью до бесконечно малых более высокого порядка, где

$$g_i = \frac{\partial^2 f(x_i; q_0)}{\partial x_i \partial q_0} A - \frac{\partial^3 f(x_i; q_0)}{\partial x_i \partial q_0^2} B_0(x).$$

Ясно, что задача оптимизации

$$\left\{ \begin{array}{l} \sum_{1 \leq i \leq n} g_i e_i \rightarrow \max \\ |e_i| \leq \Delta, \quad i = 1, 2, \dots, n, \end{array} \right. \quad (41)$$

имеет решение

$$e_i = \begin{cases} \Delta, & g_i \geq 0, \\ -\Delta, & g_i < 0, \end{cases}$$

при этом максимальное значение линейной формы есть  $\Delta \sum_{1 \leq i \leq n} |g_i|$ .

Поэтому

$$\sup_{\{e\}} |q_n^* - q_n| = \frac{\Delta}{A^2 n} \sum_{1 \leq i \leq n} |g_i|. \quad (42)$$

С целью упрощения правой части (42) воспользуемся тем, что

$$\frac{1}{n} \sum_{1 \leq i \leq n} |g_i| = \frac{|A|}{n} \sum_{1 \leq i \leq n} \left| \frac{\partial^2 f(x_i; q_0)}{\partial x \partial q_0} \right| + a \frac{|B_0(x)|}{n} \sum_{1 \leq i \leq n} \left| \frac{\partial^3 f(x_i; q_0)}{\partial x \partial q_0^2} \right|, \quad (43)$$

где  $|a| \leq 1$ . Поскольку при  $n \rightarrow \infty$

$$\frac{1}{n} \sum_{1 \leq i \leq n} \left| \frac{\partial^3 f(x_i; q_0)}{\partial x \partial q_0^2} \right| \rightarrow M \left| \frac{\partial^3 f(x_1; q_0)}{\partial x \partial q_0^2} \right| < +\infty, \quad B_0(x) \rightarrow 0$$

по вероятности, то второе слагаемое в (43) сходится к 0, а первое в силу закона больших чисел с учетом (39) сходится к  $CA^2$ , где  $C$  определено в (40). Теорема 4 доказана.

**Оценки**

**метода**

**моментов.**

Пусть

$g : R^k \rightarrow R^1, \quad h_j : R^1 \rightarrow R^1, j = 1, 2, \dots, k,$  - некоторые функции.

Рассмотрим аналоги выборочных моментов

$$m_j = \frac{1}{n} \sum_{1 \leq i \leq n} h_j(x_i), j = 1, 2, \dots, k.$$

Оценки метода моментов имеют вид

$$\hat{q}_n(x) = g(m_1, m_2, \dots, m_k)$$

(функции  $g$  и  $h_j$  должны удовлетворять некоторым дополнительным условиям [ , с.80], которые здесь не приводим). Очевидно, что

$$\begin{aligned} \hat{q}_n(y) - \hat{q}_n(x) &= \sum_{1 \leq j \leq k} \frac{\partial g}{\partial m_j} (m_j(y) - m_j(x)), \\ m_j(y) - m_j(x) &= \frac{1}{n} \sum_{1 \leq i \leq n} \frac{dh_j(x_i)}{dx_i} e_i, \quad j = 1, 2, \dots, k, \end{aligned} \quad (44)$$

с точностью до бесконечно малых более высокого порядка, а потому с той же точностью

$$\hat{q}_n(y) - \hat{q}_n(x) = \frac{1}{n} \sum_{1 \leq i \leq n} \left( \sum_{1 \leq j \leq k} \frac{\partial g}{\partial m_j} \frac{dh_j(x_i)}{dx_i} \right) e_i. \quad (45)$$

**Теорема 5.** Пусть при  $q = q_0$  существуют математические ожидания

$$M_j = Mm_j = Mh_j(x_1), \quad M \left( \frac{dh_j(x_1)}{dx_1} \right), \quad j = 1, 2, \dots, n,$$

функция  $g$  дважды непрерывно дифференцируема в некоторой окрестности точки  $(M_1, M_2, \dots, M_k)$ . Пусть существует функция  $t : R^1 \rightarrow R^1$  такая, что

$$\sup_{|x-y| \leq \Delta} \left| h_j(y) - h_j(x) - \frac{dh_j(x)}{dx} (y-x) \right| \leq t(x) \Delta^2, \quad j = 1, 2, \dots, k, \quad (46)$$

причем  $Mt(x_1)$  существует. Тогда

$$\sup_{\{e\}} |\hat{q}_n(y) - \hat{q}_n(x)| = C_1 \Delta$$

с точностью до бесконечно малых более высокого порядка, причем

$$C_1 = M \left| \sum_{1 \leq j \leq k} \frac{\partial g(M_1, M_2, \dots, M_k)}{\partial m_j} \frac{dh_j(x_1)}{dx_1} \right|.$$

*Доказательство* теоремы 5 сводится к обоснованию проведенных ранее рассуждений, позволивших получить формулу (45). В условиях теоремы 5 собраны предположения, достаточные для такого обоснования. Так, условие (46) дает возможность

обосновать соотношения (44); существование  $M\left(\frac{dh_j(x_1)}{dx_1}\right)$  обеспечивает существование  $C_1$ , и т.д. Завершает доказательство ссылка на решение задачи оптимизации (41) и применение закона больших чисел.

Полученные в теоремах 4 и 5 нотны оценок минимального контраста и метода моментов, асимптотические дисперсии этих оценок (см. теорему 2 и [40] соответственно) позволяют находить рациональные объемы выборок, строить доверительные интервалы с учетом погрешностей измерений, а также сравнивать оценки по среднему квадрату ошибки (36). Подобное сравнение было проведено для оценок максимального правдоподобия и метода моментов параметров гамма-распределения. Установлено, что классический вывод о преимуществе оценок максимального правдоподобия [33, с.99-100] неверен в случае  $\Delta > 0$ .

### **2.3.6. Интервальные данные в задачах проверки гипотез**

С позиций статистики интервальных данных целесообразно изучить все практически используемые процедуры прикладной математической статистики, установить соответствующие нотны и рациональные объемы выборок. Это позволит устранить разрыв между математическими схемами прикладной статистики и реальностью влияния погрешностей наблюдений на свойства статистических процедур. Статистика интервальных данных – часть теории устойчивых статистических процедур, развитой в монографии [3]. Часть, более адекватная реальной статистической практике, чем некоторые другие постановки, например, с засорением нормального распределения большими выбросами.

Рассмотрим подходы статистики интервальных данных в задачах проверки статистических гипотез. Пусть принятие решения

основано на сравнении рассчитанного по выборке значения статистики критерия  $f = f(y_1, y_2, \dots, y_n)$  с граничным значением  $C$ : если  $f > C$ , то гипотеза отвергается, если же  $f \leq C$ , то принимается. С учетом погрешностей измерений выборочное значение статистики критерия может принимать любое значение в интервале  $[f(y) - N_f(y); f(y) + N_f(y)]$ . Это означает, что «истинное» значение порога, соответствующее реально используемому критерию, находится между  $C - N_f(y)$  и  $C + N_f(y)$ , а потому уровень значимости описанного правила (критерия) лежит между  $1 - P(C + N_f(y))$  и  $1 - P(C - N_f(y))$ , где  $P(Z) = P(f < Z)$ .

**Пример 1.** Пусть  $x_1, x_2, \dots, x_n$  - выборка из нормального распределения с математическим ожиданием  $a$  и единичной дисперсией. Необходимо проверить гипотезу  $H_0: a = 0$  при альтернативе  $H_1: a \neq 0$ .

Как известно из любого учебного курса математической статистики, следует использовать статистику  $f = \sqrt{n} |\bar{y}|$  и порог  $C = \Phi(1 - a/2)$ , где  $a$  - уровень значимости,  $\Phi(\cdot)$  - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. В частности,  $C = 1,96$  при  $a = 0,05$ .

При ограничениях (1) на абсолютную погрешность  $N_f(y) = \sqrt{n}\Delta$ . Например, если  $\Delta = 0,1$ , а  $n = 100$ , то  $N_f(y) = 1,0$ . Это означает, что истинное значение порога лежит между 0,96 и 2,96, а истинный уровень значимости - между 0,003 и 0,34. Можно сделать и другой вывод: нулевую гипотезу  $H_0$  допустимо отклонить на уровне значимости 0,05 лишь тогда, когда  $f > 2,96$ .

Если же  $n = 400$  при  $\Delta = 0,1$ , то  $N_f(y) = 2,0$  и  $C - N_f(y) = -0,04$ , в то время как  $C + N_f(y) = 3,96$ . Таким образом, даже в случае  $x = 0$

гипотеза  $H_0$  может быть отвергнута только из-за погрешностей измерений результатов наблюдений.

Вернемся к общему случаю проверки гипотез. С учетом погрешностей измерений граничное значение  $C_a$  в статистике интервальных данных целесообразно заменить на  $C_a + N_f(y)$ . Такая замена дает гарантию, что вероятность отклонения нулевой гипотезы  $H_0$ , когда она верна, не более  $a$ . При проверке гипотез аналогом статистической погрешности, рассмотренной выше в задачах оценивания, является  $C_a$ . Суммарная погрешность имеет вид  $C_a + N_f(y)$ . Исходя из принципа уравнивания погрешностей [3], целесообразно определять рациональный объем выборки из условия

$$C_a = N_f(y).$$

Если  $f = |f_I|$ , где  $f_I$  при справедливости  $H_0$  имеет асимптотически нормальное распределение с математическим ожиданием 0 и дисперсией  $s^2/n$ , то

$$C_a = u \left( 1 - \frac{a}{2} \right) \frac{s}{\sqrt{n}} \quad (47)$$

при больших  $n$ , где  $u(1-a/2)$  - квантиль порядка  $1-a/2$  стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из (47) вытекает, что в рассматриваемом случае

$$n_{rat} = \left[ \frac{u(1-a/2)s}{N_f(y)} \right]^2.$$

В условиях примера 1  $f_1 = \bar{y}$  и

$$n_{rat} = \frac{3,84}{\Delta^2} = 384.$$

**Пример 2.** Рассмотрим статистику одновыборочного критерия Стьюдента

$$t = \sqrt{n} \frac{\bar{y}}{s(y)} = \frac{\sqrt{n}}{v},$$

где  $v$  – выборочный коэффициент вариации. Тогда с точностью до бесконечно малых более высокого порядка нотна для  $t$  имеет вид

$$N_t(y) = \frac{\sqrt{n}}{v^2} N_v(y),$$

где  $N_v(y)$  – рассмотренная ранее нотна для выборочного коэффициента вариации. Поскольку распределение статистики Стьюдента  $t$  сходится к стандартному нормальному, то небольшое изменение предыдущих рассуждений дает

$$n_{rat} = \frac{v^4 u^2 (1 - \alpha/2)}{N_v^2(y)}.$$

**Пример 3.** Рассмотрим двухвыборочный критерий Смирнова, предназначенный для проверки однородности (совпадения) функций распределения двух независимых выборок [41]. Статистика этого критерия имеет вид

$$D_{mn} = \sup_x |F_m(x) - G_n(x)|,$$

где  $F_m(x)$  – эмпирическая функция распределения, построенная по первой выборке объема  $m$ , извлеченной из генеральной совокупности с функцией распределения  $F(x)$ , а  $G_n(x)$  – эмпирическая функция распределения, построенная по второй выборке объема  $n$ , извлеченной из генеральной совокупности с функцией распределения  $G(x)$ . Нулевая гипотеза имеет вид  $H_0 : F(x) \equiv G(x)$ , альтернативная состоит в ее отрицании:

$H_1 : F(x) \neq G(x)$  при некотором  $x$ . Значение статистики сравнивают с порогом  $D(\alpha, m, n)$ , зависящим от уровня значимости  $\alpha$  и объемов выборок  $m$  и  $n$ . Если значение статистики не превосходит порога, то принимают нулевую гипотезу, если больше порога – альтернативную. Пороговые значения  $D(\alpha, m, n)$  берут из таблиц [42]. Описанный критерий иногда неправильно называют критерием Колмогорова-Смирнова. История вопроса описана в [43].

При ограничениях (1) на абсолютные погрешности и справедливости нулевой гипотезы  $H_0 : F(x) \equiv G(x)$  нотна имеет вид (при больших объемах выборок)

$$N_D = \sup_x |F(x + \Delta) - F(x - \Delta)|.$$

Если  $F(x)=G(x)=x$  при  $0 \leq x \leq 1$ , то  $N_D = 2\Delta$ . С помощью условия  $C_a = N_f(y)$  при уровне значимости  $a = 0,05$  и достаточно больших объемах выборок (т.е. используя асимптотическое выражение для порога согласно [42]) получаем, что выборки имеет смысл увеличивать, если

$$\frac{mn}{m+n} \leq \frac{0,46}{\Delta^2}.$$

Правая часть этой формулы при  $\Delta = 0,1$  равна 46. Если  $m = n$ , то последнее неравенство переходит в  $n \leq 92$ .

Теоретические результаты в области статистических методов входят в практику через алгоритмы расчетов, воплощенные в программные средства (пакеты программ, диалоговые системы). Ввод данных в современной статистической программной системе должен содержать запросы о погрешностях результатов измерений. На основе ответов на эти запросы вычисляются нотны рассматриваемых статистик, а затем – доверительные интервалы при оценивании, разброс уровней значимости при проверке гипотез, рациональные объемы выборок. Необходимо использовать систему алгоритмов и программ статистики интервальных данных, «параллельную» подобным системам для классической математической статистики.

### **2.3.7. Асимптотический линейный регрессионный**

#### **анализ**

#### **для интервальных данных**

Перейдем к многомерному статистическому анализу. Сначала с позиций асимптотической математической статистики интервальных данных рассмотрим оценки метода наименьших квадратов (МНК).

Статистическое исследование зависимостей - одна из наиболее важных задач, которые возникают в различных областях науки и техники. Под словами "исследование зависимостей" имеется в виду выявление и описание существующей связи между исследуемыми переменными на основании результатов статистических наблюдений. К методам исследования зависимостей относятся регрессионный анализ, многомерное шкалирование, идентификация параметров динамических объектов, факторный анализ, дисперсионный анализ, корреляционный анализ и др. Однако многие реальные ситуации характеризуются наличием данных интервального типа, причем известны допустимые границы погрешностей (например, из технических паспортов средств измерения).

Если какая-либо группа объектов характеризуется переменными  $X_1, X_2, \dots, X_m$  и проведен эксперимент, состоящий из  $n$  опытов, где в каждом опыте эти переменные измеряются один раз, то экспериментатор получает набор чисел:  $X_{1j}, X_{2j}, \dots, X_{mj}$  ( $j = 1, \dots, n$ ).

Однако процесс измерения, какой бы физической природы он ни был, обычно не дает однозначный результат. Реально результатом измерения какой-либо величины  $X$  являются два числа:  $X_H$  — нижняя граница и  $X_B$  — верхняя граница. Причем  $X_{ИСТ} \hat{I} [X_H, X_B]$ , где  $X_{ИСТ}$  - истинное значение измеряемой величины. Результат измерения можно записать как  $X: [X_H, X_B]$ . Интервальное число  $X$  может быть представлено другим способом, а именно,  $X: [X_m, \Delta_x]$ , где  $X_H = X_m - \Delta_x$ ,  $X_B = X_m + \Delta_x$ . Здесь  $X_m$  - центр интервала (как

правило, не совпадающий с  $X_{ИСТ}$ ), а  $\Delta_x$  - максимально возможная погрешность измерения.

### **Метод наименьших квадратов для интервальных данных.**

Пусть математическая модель задана следующим образом:

$$y = Q(x, b) + \varepsilon,$$

где  $x = (x_1, x_2, \dots, x_m)$  - вектор влияющих переменных (факторов), поддающихся измерению;  $b = (b_1, b_2, \dots, b_r)$  - вектор оцениваемых параметров модели;  $y$  - отклик модели (скаляр);  $Q(x, b)$  - скалярная функция векторов  $x$  и  $b$ ; наконец,  $\varepsilon$  - случайная ошибка (невязка, погрешность).

Пусть проведено  $n$  опытов, причем в каждом опыте измерены (один раз) значения отклика ( $y$ ) и вектора факторов ( $x$ ). Результаты измерений могут быть представлены в следующем виде:

$$X = \{ x_{ij}; i = 1, n; j = 1, m \}, Y = (y_1, y_2, \dots, y_n), E = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n),$$

где  $X$  - матрица значений измеренного вектора ( $x$ ) в  $n$  опытах;  $Y$  - вектор значений измеренного отклика в  $n$  опытах;  $E$  - вектор случайных ошибок. Тогда выполняется матричное соотношение:

$$Y = Q(X, b) + E,$$

где  $Q(X, b) = (Q(x_1, b), Q(x_2, b), \dots, Q(x_n, b))^T$ , причем  $x_1, x_2, \dots, x_n$  -  $m$ -мерные вектора, которые составляют матрицу  $X = (x_1, x_2, \dots, x_n)^T$ .

Введем меру близости  $d(Y, Q)$  между векторами  $Y$  и  $Q$ . В МНК в качестве  $d(Y, Q)$  берется квадратичная форма взвешенных квадратов  $\varepsilon_i^2$  невязок  $\varepsilon_i = y_i - Q(x_i, b)$ , т.е.

$$d(Y, Q) = [Y - Q(X, b)]^T W [Y - Q(X, b)],$$

где  $W = \{w_{ij}, i, j = 1, \dots, n\}$  - матрица весов, не зависящая от  $b$ . Тогда в качестве оценки  $b$  можно выбрать такое  $b^*$ , при котором мера близости  $d(Y, Q)$  принимает минимальное значение, т.е.

$$b^* = \{b : d(Y, Q) \rightarrow \min_{\{b\}}\}.$$

В общем случае решение этой экстремальной задачи может быть не единственным. Поэтому в дальнейшем будем иметь в виду одно из этих решений. Оно может быть выражено в виде  $b^* = f(X, Y)$ , где  $f(X, Y) = (f_1(X, Y), f_2(X, Y), \dots, f_m(X, Y))^T$ , причем  $f_i(X, Y)$  непрерывны и дифференцируемы по  $(X, Y) \in Z$ , где  $Z$  - область определения функции  $f(X, Y)$ . Эти свойства функции  $f(X, Y)$  дают возможность использовать подходы статистики интервальных данных.

Преимущество метода наименьших квадратов заключается в сравнительной простоте и универсальности вычислительных процедур. Однако не всегда оценка МНК является состоятельной (при функции  $Q(X, b)$ , не являющейся линейной по векторному параметру  $b$ ), что ограничивает его применение на практике.

Важным частным случаем является линейный МНК, когда  $Q(x, b)$  есть линейная функция от  $b$ :

$$y = b_0 x_0 + b_1 x_1 + \dots + b_m x_m + \varepsilon = b x^T + \varepsilon,$$

где, возможно,  $x_0 = 1$ , а  $b_0$  - свободный член линейной комбинации. Как известно, в этом случае МНК-оценка имеет вид:

$$b^* = (X^T W X)^{-1} X^T W Y.$$

Если матрица  $X^T W X$  не вырождена, то эта оценка является единственной. Если матрица весов  $W$  единичная, то

$$b^* = (X^T X)^{-1} X^T Y.$$

Пусть выполняются следующие предположения относительно распределения ошибок  $\varepsilon_i$ :

- ошибки  $\varepsilon_i$  имеют нулевые математические ожидания  $M\{\varepsilon_i\} = 0$ ,
- результаты наблюдений имеют одинаковую дисперсию  $D\{\varepsilon_i\} = \sigma^2$ ,
- ошибки наблюдений некоррелированы, т.е.  $cov\{\varepsilon_i, \varepsilon_j\} = 0$ .

Тогда, как известно, оценки МНК являются наилучшими линейными оценками, т.е. состоятельными и несмещенными оценками, которые представляют собой линейные функции результатов наблюдений и обладают минимальными дисперсиями

среди множества всех линейных несмещенных оценок. Далее именно этот наиболее практически важный частный случай рассмотрим более подробно.

Как и в других постановках асимптотической математической статистики интервальных данных, при использовании МНК измеренные величины отличаются от истинных значений из-за наличия погрешностей измерения. Запишем истинные данные в следующей форме:

$$X_R = \{ x_{ij}^R ; i = \overline{1, n} ; j = \overline{1, m} \}, Y_R = ( y_1^R, y_2^R, \dots, y_n^R ),$$

где  $R$  - индекс, указывающий на то, что значение истинное.

Истинные и измеренные данные связаны следующим образом:

$$X = X_R + \Delta X, Y = Y_R + \Delta Y,$$

где  $\Delta X = \{ \Delta x_{ij} ; i = \overline{1, n} ; j = \overline{1, m} \}, \Delta Y = (\Delta y_1, \Delta y_2, \dots, \Delta y_n)$ . Предположим, что погрешности измерения отвечают граничным условиям

$$|\Delta x_{ij}| \leq \Delta^x \quad (i = 1, 2, \dots, n, j = 1, 2, \dots, m), \quad |\Delta y_i| \leq \Delta^y \quad (i = 1, 2, \dots, n), \quad (48)$$

аналогичным ограничениям (1).

Пусть множество  $W$  возможных значений  $(X_R, Y_R)$  входит в  $Z$  - область определения функции  $f(X, Y)$ . Рассмотрим  $b^{*R}$  - оценку МНК, рассчитанную по истинным значениям факторов и отклика, и  $b^*$  - оценку МНК, найденную по искаженным погрешностями данным. Тогда

$$\Delta b^* = b^{*R} - b^* = f(X_R, Y_R) - f(X, Y).$$

Ввести понятие *нотны* придется несколько иначе, чем это было сделано выше, поскольку оценивается не одномерный параметр, а вектор. Положим:

$$n(1) = (\sup \Delta b_1^*, \sup \Delta b_2^*, \dots, \sup \Delta b_r^*)^T, \quad n(2) = -(\inf \Delta b_1^*, \inf \Delta b_2^*, \dots, \inf \Delta b_r^*)^T.$$

Будем называть  $n(1)$  нижней *нотной*, а  $n(2)$  верхней *нотной*. Предположим, что при безграничном возрастании числа измерений  $n$ , т.е. при  $n \rightarrow \infty$ , вектора  $n(1)$ ,  $n(2)$  стремятся к постоянным значениям  $N_i(1)$ ,  $N_i(2)$  соответственно. Тогда  $N_i(1)$  будем называть

нижней асимптотической нотной, а  $N_i(2)$  - верхней асимптотической нотной.

Рассмотрим доверительное множество  $B_\alpha = B_\alpha(n, b^{*R})$  для вектора параметров  $b$ , т.е. замкнутое связное множество точек в  $r$ -мерном евклидовом пространстве такое, что  $P(b \in B_\alpha) = \alpha$ , где  $\alpha$  — доверительная вероятность, соответствующая  $B_\alpha$  ( $\alpha \approx 1$ ). Другими словами,  $B_\alpha(n, b^{*R})$  есть область рассеивания (аналог эллипсоида рассеивания) случайного вектора  $b^{*R}$  с доверительной вероятностью  $\alpha$  и числом опытов  $n$ .

Из определения верхней и нижней нотн следует, что всегда  $b^{*R} \in [b^* - n(1); b^* + n(2)]$ . В соответствии с определением нижней асимптотической нотны и верхней асимптотической нотны можно считать, что  $b^{*R} \in [b^* - N(1); b^* + N(2)]$  при достаточно большом числе наблюдений  $n$ . Этот многомерный интервал описывает  $r$ -мерный гиперпараллелепипед  $P$ .

Каким-либо образом разобьем  $P$  на  $L$  гиперпараллелепипедов. Пусть  $b_k$  - внутренняя точка  $k$ -го гиперпараллелепипеда. Учитывая свойства доверительного множества и устремляя  $L$  к бесконечности, можно утверждать, что  $P(b \in C) \geq \alpha$ , где

$$C = \lim_{L \rightarrow \infty} \bigcup_{1 \leq k \leq L} B_\alpha(n, b_k).$$

Таким образом, множество  $C$  характеризует неопределенность при оценивании вектора параметров  $b$ . Его можно назвать доверительным множеством в статистике интервальных данных.

Введем некоторую меру  $M(X)$ , характеризующую «величину» множества  $X \subseteq R^r$ . По определению меры она удовлетворяет условию: если  $X = Z \cup Y$  и  $Z \cap Y = 0$ , то  $M(X) = M(Z) + M(Y)$ . Примерами такой меры являются площадь для  $r = 2$  и объем для  $r = 3$ . Тогда:

$$M(C) = M(P) + M(F), \quad (49)$$

где  $F = C \setminus P$ . Здесь  $M(F)$  характеризует меру статистической неопределенности, в большинстве случаев она убывает при увеличении числа опытов  $n$ . В то же время  $M(P)$  характеризует меру интервальной (метрологической) неопределенности, и, как правило,  $M(P)$  стремится к некоторой постоянной величине при увеличении числа опытов  $n$ . Пусть теперь требуется найти то число опытов, при котором статистическая неопределенность составляет  $\delta$ -ю часть общей неопределенности, т.е.

$$M(F) = \delta M(C), \quad (50)$$

где  $\delta < 1$ . Тогда, подставив соотношение (50) в равенство (49) и решив уравнение относительно  $n$ , получим искомое число опытов. В асимптотической математической статистике интервальных данных оно называется "рациональным объемом выборки". При этом  $\delta$  есть "степень малости" статистической неопределенности  $M(P)$  относительно всей неопределенности. Она выбирается из практических соображений. При использовании "принципа уравнивания погрешностей" согласно [3] имеем  $\delta = 1/2$ .

#### **Метод наименьших квадратов для линейной модели.**

Рассмотрим наиболее важный для практики частный случай МНК, когда модель описывается линейным уравнением (см. выше).

Для простоты описания преобразований пронормируем переменные  $x_{ij}, y_i$  следующим образом:

$$x_{ij}^0 = (x_{ij} - \bar{x}_j) / s(x_j), \quad y_i^0 = (y_i - \bar{y}) / s(y),$$

где

$$\bar{x}_j = \frac{1}{n} \sum_{1 \leq i \leq n} x_{ij}, \quad s^2(x_j) = \frac{1}{n} \sum_{1 \leq i \leq n} (x_{ij} - \bar{x}_j)^2, \quad \bar{y} = \frac{1}{n} \sum_{1 \leq i \leq n} y_i, \quad s^2(y) = \frac{1}{n} \sum_{1 \leq i \leq n} (y_i - \bar{y})^2.$$

Тогда

$$\bar{x}_j^0 = 0, \quad s^2(x_j^0) = \frac{1}{n} \sum_{1 \leq i \leq n} (x_{ij}^0 - \bar{x}_j^0)^2 = 1, \quad \bar{y}^0 = 0, \quad s^2(y^0) = \frac{1}{n} \sum_{1 \leq i \leq n} (y_i^0 - \bar{y}^0)^2 = 1, \quad j = 1, 2, \dots, m.$$

В дальнейшем изложении будем считать, что рассматриваемые переменные пронормированы описанным

образом, и верхние индексы <sup>0</sup> опустим. Для облегчения демонстрации основных идей примем достаточно естественные предположения.

1. Для рассматриваемых переменных существуют следующие пределы:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{1 \leq i \leq n} x_{ij} x_{ik} = 0, \quad j, k = 1, 2, \dots, m.$$

2. Количество опытов  $n$  таково, что можно пользоваться асимптотическими результатами, полученными при  $n \rightarrow \infty$ .

3. Погрешности измерения удовлетворяют одному из следующих типов ограничений:

*Тип 1.* Абсолютные погрешности измерения ограничены согласно (48):

*Тип 2.* Относительные погрешности измерения ограничены:

$$|\Delta x_{ij}| \leq d_j^x |x_{ij}| \quad (i = 1, 2, \dots, n, j = 1, 2, \dots, m), \quad |\Delta y_i| \leq d_i^y |y_i| \quad (i = 1, 2, \dots, n).$$

*Тип 3.* Ограничения наложены на сумму погрешностей:

$$\sum_{j=1}^m |\Delta x_{ij}| \leq a_x \quad (i = 1, 2, \dots, n, j = 1, 2, \dots, m), \quad |\Delta y_i| \leq a_y \quad (i = 1, 2, \dots, n).$$

(поскольку все переменные отнормированы, т.е. представляют собой относительные величины, то различие в размерности исходных переменных не влияет на возможность сложения погрешностей).

Перейдем к вычислению нотны оценки МНК. Справедливо равенство:

$$\Delta b^* = b^{*R} - b^* = (X_R^T X_R)^{-1} X_R^T Y_R - (X^T X)^{-1} X^T Y = (X_R^T X_R)^{-1} X_R^T Y_R - ((X_R + \Delta X)^T (X_R + \Delta X))^{-1} (X_R + \Delta X)^T (Y_R + \Delta Y).$$

Воспользуемся следующей теоремой из теории матриц [14].

**Теорема.** Если функция  $f(\lambda)$  разлагается в степенной ряд в круге сходимости  $|\lambda - \lambda_0| < r$ , т.е.

$$f(I) = \sum_{k=0}^{\infty} a_k (I - I_0)^k,$$

то это разложение сохраняет силу, если скалярный аргумент заменить любой матрицей  $A$ , характеристические числа которой  $\lambda_k$ ,  $k = 1, \dots, n$ , лежат внутри круга сходимости.

$$(E - A)^{-1} = \sum_{P=0}^{\infty} A^P, \quad \text{если} \quad |I_k| < 1; \quad k = 1, \dots, n.$$

Из этой теоремы вытекает, что:

Легко убедиться, что:

$$((X_R + \Delta X)^T (X_R + \Delta X))^{-1} = -Z (E - \Delta \cdot Z)^{-1},$$

где  $Z = -(X_R^T X_R)^{-1}$ ,  $\Delta = X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X$ .

Это вытекает из последовательности равенств:

$$\begin{aligned} ((X_R + \Delta X)^T (X_R + \Delta X))^{-1} &= (X_R^T X_R + X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X)^{-1} = (X_R^T X_R + \Delta)^{-1} = \\ &= ((E + \Delta (X_R^T X_R)^{-1} X_R^T X_R)^{-1} = (X_R^T X_R)^{-1} (E + \Delta (X_R^T X_R)^{-1})^{-1} = -Z (E - \Delta \cdot Z)^{-1}. \end{aligned}$$

Применим приведенную выше теорему из теории матриц, полагая  $A = \Delta Z$  и принимая, что собственные числа этой матрицы удовлетворяют неравенству  $|\lambda_k| < 1$ . Тогда получим:

$$((X_R + \Delta X)^T (X_R + \Delta X))^{-1} = -Z \sum_{P=0}^{\infty} (\Delta \cdot Z)^P = (X_R^T X_R)^{-1} \sum_{P=0}^{\infty} (-\Delta \cdot (X_R^T X_R)^{-1})^P.$$

Подставив последнее соотношение в заключение упомянутой теоремы, получим:

$$\begin{aligned} \Delta b^* &= (X_R^T X_R)^{-1} X_R^T Y_R - ((X_R^T X_R)^{-1} \sum_{P=0}^{\infty} (-\Delta \cdot (X_R^T X_R)^{-1})^P) (X_R + \Delta X)^T (Y_R + \Delta Y) = \\ &= (X_R^T X_R)^{-1} X_R^T Y_R - ((X_R^T X_R)^{-1} \sum_{P=0}^{\infty} (-\Delta \cdot (X_R^T X_R)^{-1})^P) (X_R^T Y_R + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y). \end{aligned}$$

Для дальнейшего анализа понадобится вспомогательное утверждение. Исходя из предположений 1-3, докажем, что:

$$(X_R^T X_R)^{-1} \approx \frac{1}{n} E.$$

*Доказательство.* Справедливо равенство

$$X_R^T X_R = n \cdot \begin{pmatrix} \hat{D}(x_1) & \dots & \hat{\text{cov}}(x_1, x_m) \\ \dots & \dots & \dots \\ \hat{\text{cov}}(x_1, x_m) & \dots & \hat{D}(x_m) \end{pmatrix} = n \cdot \hat{\text{cov}}(x),$$

где  $\hat{D}(x_i)$ ,  $\hat{\text{cov}}(x_i, x_j)$  - состоятельные и несмещенные оценки дисперсий и коэффициентов ковариации, т.е.

$$\hat{D}(x_i) = D(x_i) + o(1/n), \quad \hat{\text{cov}}(x_i, x_j) = \text{cov}(x_i, x_j) + o(1/n),$$

тогда

$$X_R^T X_R = n \cdot \hat{\text{cov}}(x) = n \cdot (\text{cov}(x_i, x_j) + o(1/n)),$$

где

$$o(1/n) = \{a_{ij} = o(1/n)\} \quad (i = \overline{1, n}, j = \overline{1, m}).$$

Другими словами, каждый элемент матрицы, обозначенной как  $o(1/n)$ , есть бесконечно малая величина порядка  $1/n$ . Для рассматриваемого случая  $\text{cov}(x) = E$ , поэтому

$$X_R^T X_R = n \cdot \hat{\text{cov}}(x) = n \cdot (E + o(1/n)).$$

Предположим, что  $n$  достаточно велико и можно считать, что собственные числа матрицы  $o(1/n)$  меньше единицы по модулю, тогда

$$(X_R^T X_R)^{-1} = \frac{1}{n} \cdot (E + o(1/n))^{-1} \approx \frac{1}{n} (E + o(1/n)) = \frac{1}{n} E + o(1/n^2) \approx \frac{1}{n} E,$$

что и требовалось доказать.

Подставим доказанное асимптотическое соотношение в формулу для приращения  $b^*$ , получим

$$\begin{aligned}
\Delta b^* &= b^{*R} - \frac{1}{n} \sum_{P=0}^{\infty} \left(-\Delta \cdot \frac{1}{n}\right)^P (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y) = \\
&= b^{*R} - \frac{1}{n} \sum_{P=0}^{\infty} \left(-\left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right) \cdot \left(\frac{1}{n}\right)^P (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y)\right) = \\
&= b^{*R} - \frac{1}{n} \left(E - \left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right) \frac{1}{n} + \left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right)^2 \left(\frac{1}{n}\right)^2\right) \cdot \\
&\cdot (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y).
\end{aligned}$$

Выразим  $\Delta b^*$  относительно приращений  $\Delta X$ ,  $\Delta Y$  до 2-го порядка

$$\begin{aligned}
\Delta b^* &= b^{*R} - \frac{1}{n} \left(E - \left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right) \frac{1}{n} + \left(X_R^T \Delta X X_R^T \Delta X + \Delta X^T X_R \Delta X^T X_R + \right. \right. \\
&\left. \left. + \Delta X^T X_R X_R^T \Delta X + X_R^T \Delta X \Delta X^T X_R\right) \left(\frac{1}{n}\right)^2\right) \cdot (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y); \\
\Delta b^* &= b^{*R} - \frac{1}{n} \left(E - \left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right) \frac{1}{n}\right) \cdot (nb^{*R} + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y); \\
\Delta b^* &= \frac{1}{n} \left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right) b^{*R} - \frac{1}{n} \left(\Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y\right) = \\
&= \frac{1}{n} \left[\left(X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X\right) b^{*R} - \left(\Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y\right)\right].
\end{aligned}$$

Перейдем от матричной к скалярной форме, опуская индекс ( $R$ ):

$$\begin{aligned}
\Delta b_k^* &= \frac{1}{n} \left\{ \sum_j^m \sum_i^n (x_{ik} \Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \sum_i^n (\Delta x_{ik} y_i + x_{ik} \Delta y_i) \right\}; \\
\Delta b_k^* &= \frac{1}{n} \left\{ 2 \sum_i^n x_{ik} \Delta x_{ik} b_k^* + \sum_{j \neq k}^m \sum_i^n (x_{ik} \Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \sum_i^n (\Delta x_{ik} y_i + x_{ik} \Delta y_i) \right\} = \\
&= \frac{1}{n} \left\{ 2 \sum_i^n x_{ik} \Delta x_{ik} b_k^* + \sum_{j \neq k}^m \sum_i^n [(x_{ik} \Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \frac{1}{m-1} \Delta x_{ik} y_i] - \sum_i^n x_{ik} \Delta y_i \right\} = \\
&= \frac{1}{n} \left\{ \sum_{j \neq k}^m \sum_i^n \left[ \frac{2}{m-1} x_{ik} \Delta x_{ik} b_k^* + (x_{ik} \Delta x_{ij} + \Delta x_{ik} x_{ij}) b_j^* - \frac{1}{m-1} \Delta x_{ik} y_i \right] - \sum_i^n x_{ik} \Delta y_i \right\} = \\
&= \frac{1}{n} \left\{ \sum_{j \neq k}^m \sum_i^n \left[ \left( \frac{2}{m-1} x_{ik} b_k^* + x_{ij} b_j^* - \frac{1}{m-1} y_i \right) \Delta x_{ik} - x_{ik} b_j^* \Delta x_{ij} \right] - \sum_i^n x_{ik} \Delta y_i \right\}
\end{aligned}$$

Будем искать  $\max(|\Delta b_k^*|)$  по  $\Delta x_{ij}$  и  $\Delta y_i$  ( $i=1, \dots, n$ ;  $j=1, \dots, m$ ).

Для этого рассмотрим все три ранее введенных типа ограничений на ошибки измерения.

*Тип 1* (абсолютные погрешности измерения ограничены).

Тогда:

$$\max_{\Delta x, \Delta y} (|\Delta b_k^*|) = \frac{1}{n} \left\{ \sum_{j \neq k}^m \sum_i^n \left[ \left( \frac{2}{m-1} x_{ik} b_k^* + x_{ij} b_j^* - \frac{1}{m-1} y_i \right) |\Delta x_k^*| + |x_{ik} b_j^*| |\Delta x_j^*| \right] - \sum_i^n x_{ik} |\Delta y_i^*| \right\}.$$

*Тун 2* (относительные погрешности измерения ограничены).

Аналогично получим:

$$\sum_{j=1}^m |\Delta x_{ij}| < a_x \quad (i = 1, 2, \dots, n, j = 1, 2, \dots, m), \quad |\Delta y_i| < a_y \quad (i = 1, 2, \dots, n).$$

*Тун 3* (ограничения наложены на сумму погрешностей).

Предположим, что  $|\Delta b_k^*|$  достигает максимального значения при таких значениях погрешностей  $\Delta x_{ij}$  и  $\Delta y_i$ , которые мы обозначим как:

$$\{\Delta x_{ij}^*, \quad i = \overline{1, 2, \dots, n}; j = \overline{1, 2, \dots, m}\}, \quad \{\Delta y_i^*, \quad i = \overline{1, 2, \dots, n}\}.$$

тогда:

$$\max_{\Delta x, \Delta y} (|\Delta b_k^*|) = \frac{1}{n} \left\{ \sum_{j \neq k}^m \sum_i^n \left[ \left( \frac{2}{m-1} x_{ik} b_k^* + x_{ij} b_j^* - \frac{1}{m-1} y_i \right) x_{ik}^* + x_{ik} b_j^* x_{ij}^* \right] - \sum_i^n x_{ik} y_i^* \right\}.$$

Ввиду линейности последнего выражения и выполнения ограничения типа 3:

$$\max_{\Delta x, \Delta y} (|\Delta b_k^*|) = \frac{1}{n} \left\{ \sum_{j \neq k}^m \sum_i^n \left[ \left| \frac{2}{m-1} x_{ik} b_k^* + x_{ij} b_j^* - \frac{1}{m-1} y_i \right| \cdot |\Delta x_{ik}^*| + |x_{ik} b_j^*| \cdot |\Delta x_{ij}^*| \right] - \sum_i^n |x_{ik}| \cdot |\Delta y_i^*| \right\},$$

$$\sum_j^m |\Delta x_{ij}^*| = a_x \quad (j = \overline{1, 2, \dots, m}), \quad |\Delta y_i^*| = a_y.$$

Для простоты записей выкладок сделаем следующие замены:

$$|\Delta x_{ij}| = a_{ij} \geq 0, \quad C_k = n \sum_i^n |x_{ik}| \cdot |\Delta y_i^*| \geq 0,$$

$$K_i^k = \sum_{j \neq k}^m \left| \frac{2}{m-1} x_{ik} b_k^* + x_{ij} b_j^* - \frac{1}{m-1} y_i \right| \geq 0,$$

$$|x_{ik} b_j^*| = R_{ij}^k \geq 0.$$

Теперь для достижения поставленной цели можно сформулировать следующую задачу, которая разделяется на  $m$  типовых задач оптимизации:

$$f_k(\{a_{ij}\}) \rightarrow \max_{a_{ij}} \quad (i=1,2,\dots,n; j=1,2,\dots,m; k=1,2,\dots,m),$$

где

$$f_k(\{a_{ij}\}) = \frac{1}{n} \left\{ \sum_i^n K_i^k a_{ik} + \sum_{j \neq m}^m \sum_i^n R_{ij}^k a_{ij} \right\} + C_k,$$

при ограничениях

$$\sum_j^m a_{ij} = a_x \quad (j=1,2,\dots,m).$$

Перепишем минимизируемые функции в следующем виде:

$$f_k = \frac{1}{n} \sum_i^n (K_i^k a_{ik} + \sum_{j \neq m}^m R_{ij}^k a_{ij}) + C_k = \frac{1}{n} \sum_i^n f_i^k + C_k.$$

Очевидно, что  $f_i^k > 0$ .

Легко видеть, что

$$n \cdot \max_{a_{ij}}(f_k) = \max_{a_{i1}}(f_1^k) + \max_{a_{i2}}(f_2^k) + \dots + \max_{a_{in}}(f_n^k) + C_k = \sum_i^n \max_{a_{ii}}(f_i^k) + C_k,$$

где  $i=1,2,\dots,n; j=1,2,\dots,m$ .

Следовательно, необходимо решить  $nm$  задач

$$\{f_i^k\} \rightarrow \max_{a_{ij}} \quad (i=1,2,\dots,n; j=1,2,\dots,m; k=1,2,\dots,m)$$

при ограничениях "типа равенства":

$$\sum_j^m a_{ij} = a_x \quad (i=1,2,\dots,n),$$

$$\text{где} \quad f_i^k = K_i^k a_{ik} + \sum_{j \neq m}^m R_{ij}^k a_{ij} = \sum_j^m S_{ij}^k a_{ij},$$

$$\text{причем} \quad S_{ij}^k = \begin{cases} K_i^k, & \text{если } j = k, \\ R_{ij}^k, & \text{если } j \neq k. \end{cases}$$

Сформулирована типовая задача поиска экстремума функции.

Она легко решается. Поскольку

$$\max_{a_{ij}} (f_i^k) = \max_j (S_{ij}^k) \cdot a_x,$$

то максимальное отклонение МНК-оценки  $k$ -ого параметра равно

$$\max_{\Delta X, \Delta Y} (|\Delta \hat{b}_k|) = \max_{a_{ij}} (f_k) = \frac{1}{n} a_x \sum_i^n \max_j (S_{ij}^k) + \frac{1}{n} C_k, \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, m).$$

Кроме рассмотренных выше трех видов ограничений на погрешности могут представлять интерес и другие, но для демонстрации типовых результатов ограничимся только этими тремя видами.

**Оценивание линейной корреляционной связи.** В качестве примера рассмотрим оценивание линейной корреляционной связи случайных величин  $y$  и  $x_1, x_2, \dots, x_m$  с нулевыми математическими ожиданиями. Пусть эта связь описывается соотношением:

$$y = \sum_{j=1}^m b_j x_j + e,$$

где  $b_1, b_2, \dots, b_m$  - постоянные, а случайная величина  $e$  некоррелирована с  $x_1, x_2, \dots, x_m$ . Допустим, необходимо оценить неизвестные параметры  $b_1, b_2, \dots, b_m$  по серии независимых испытаний:

$$y_i = \sum_{j=1}^m b_j x_{ij} + e_i, \quad (i = 1, 2, \dots, n).$$

Здесь при каждом  $i = 1, 2, \dots, n$  имеем новую независимую реализацию рассматриваемых случайных величин. В этой частной схеме оценки наименьших квадратов  $b_1^{*R}, b_2^{*R}, \dots, b_m^{*R}$  параметров  $b_1, b_2, \dots, b_m$  являются, как известно, состоятельными [45].

Пусть величины  $x_1, x_2, \dots, x_m$  в дополнение к попарной независимости имеют единичные дисперсии. Тогда из закона больших чисел [45] следует существование следующих пределов (ср. предположение 1 выше):

$$\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_i^n x_{ij}^R \right\} = M \{x_j\} = 0 \quad (j = \overline{1, m}),$$

$$\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_i^n (x_{ij}^R - M \{x_j\})^2 \right\} = D \{x_j\} = 1 \quad (j = \overline{1, m}),$$

$$\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_i^n (x_{ij}^R - M \{x_j\})(x_{ik}^R - M \{x_k\}) \right\} = 0 \quad (j, k = \overline{1, m}),$$

$$\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_i^n y_i^R \right\} = M \{y\} = b_1 M \{x_1\} + \dots + b_m M \{x_m\} + M \{e\} = 0,$$

$$\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_i^n (y_i^R - M \{y\})^2 \right\} = D \{y\} = b_1^2 + \dots + b_m^2 + \sigma^2,$$

где  $\sigma$  - среднее квадратическое отклонение случайной величины  $e$ .

Пусть измерения производятся с погрешностями, удовлетворяющими ограничениям типа 1, тогда максимальное приращение величины  $|\Delta b^*_{*k}|$ , как показано выше, равно:

$$\max_{\Delta x, \Delta y} (|\Delta b^*_{*k}|) = \frac{1}{n} \left\{ \sum_{j \neq k}^m \sum_i^n \left[ \frac{2}{m-1} x_{ik}^R b^*_{*k} + x_{ij}^R b^*_{*j} - \frac{1}{m-1} y_i^R \cdot \Delta x_k + |x_{ik}^R b^*_{*j}| \cdot \Delta x_j \right] + \sum_i^n |x_{ik}^R| \cdot \Delta y \right\}.$$

Перейдем к предельному случаю и выпишем выражение для нотны:

$$\begin{aligned} N_k &= \lim_{n \rightarrow \infty} \left\{ \max_{\Delta x, \Delta y} (|\Delta b^*_{*k}|) \right\} = \\ &= \sum_{j \neq k}^m \left[ M \left\{ \left| \frac{2}{m-1} x_k b_k + x_j b_j - \frac{1}{m-1} y \right| \right\} \cdot \Delta x_k + M \{ |x_k b_j| \} \cdot \Delta x_j + M \{ |x_k| \} \cdot \Delta y \right]. \end{aligned}$$

В качестве примера рассмотрим случай  $m = 2$ . Тогда

$$N_1 = M \{ |2x_1 b_1 + x_2 b_2 - y| \} \Delta x_1 + M \{ b_2 x_1 \} \Delta x_2 + M \{ |x_1| \} \Delta y,$$

$$N_2 = M \{ |2x_2 b_2 + x_1 b_1 - y| \} \Delta x_2 + M \{ b_1 x_2 \} \Delta x_1 + M \{ |x_2| \} \Delta y.$$

Приведенное выше выражение для максимального приращения метрологической погрешности не может быть использована в случае  $m = 1$ . Для  $m = 1$  выведем выражение для нотны, исходя из соотношения:

$$\Delta b^*_{*k} = \frac{1}{n} \left\{ \sum_j^m \sum_i^n (x_{ik} \Delta x_{ij} + \Delta x_{ik} x_{ij}), \quad b^*_{*j} - \sum_i^n (\Delta x_{ik} y_i + x_{ik} \Delta y_i) \right\}.$$

Подставив  $m = 1$ , получим:

$$\Delta b^* = \frac{1}{n} \left\{ \sum_i^n (2x_i \Delta x_i) b^* - \sum_i^n (\Delta x_i y_i + x_i \Delta y_i) \right\} = \frac{1}{n} \left\{ \sum_i^n ((2x_i b^* - y_i) \Delta x_i + x_i \Delta y_i) \right\}.$$

Следовательно, *нотна* выглядит так:

$$N_f = M\{2xb^* - y\} \Delta x + M\{x\} \Delta y.$$

Для нахождения рационального объема выборки необходимо сделать следующее.

*Этап 1.* Выразить зависимость размеров и меры области рассеивания  $B_\alpha(n, b)$  от числа опытов  $n$  (см. выше).

*Этап 2.* Ввести меру неопределенности и записать соотношение между статистической и интервальной неопределенностями.

*Этап 3.* По результатам этапов 1 и 2 получить выражение для рационального объема выборки.

Для выполнения этапа 1 определим область рассеивания следующим образом. Пусть доверительным множеством  $B_\alpha(n, b)$  является  $m$ -мерный куб со сторонами длиной  $2K$ , для которого

$$P(b \in B_\alpha(n, b^{*R})) = \alpha.$$

Исследуем случайный вектор  $b^*$  и

$$\begin{aligned} b^{*R} &= (X_R^T X_R)^{-1} X_R^T Y_R = (X_R^T X_R)^{-1} X_R^T (X_R b + e) = \\ &= (X_R^T X_R)^{-1} X_R^T X_R b + (X_R^T X_R)^{-1} X_R^T e = b + (X_R^T X_R)^{-1} X_R^T e. \end{aligned}$$

Как известно, если элементы матрицы  $A = \{a_{ij}\}$  -случайные, т.е.  $A$  – случайная матрица, то ее математическим ожиданием является матрица, составленная из математических ожиданий ее элементов, т.е.  $M\{A\} = \{M\{a_{ij}\}\}$ .

*Утверждение 1.* Пусть  $A = \{a_{ij}\}$  и  $B = \{b_{ij}\}$  - случайные матрицы порядка  $(m \times n)$  и  $(n \times r)$  соответственно, причем любая пара их элементов  $(a_{ij}, b_{kl})$  состоит из независимых случайных величин. Тогда математическое ожидание произведения матриц равно произведению математических ожиданий сомножителей, т.е.  $M\{AB\} = M\{A\} M\{B\}$ .

*Доказательство.* На основании определения математического ожидания матрицы заключаем, что

$$A \cdot B = \left\{ \sum_k^n a_{ik} \cdot b_{kj} \right\} \rightarrow M\{A \cdot B\} = \left\{ M \left\{ \sum_k^n a_{ik} \cdot b_{kj} \right\} \right\} = \left\{ \sum_k^n M\{a_{ik} \cdot b_{kj}\} \right\},$$

но так как случайные величины  $a_{ik}$ ,  $b_{kj}$  независимы, то

$$M\{A \cdot B\} = \left\{ \sum_k^n M\{a_{ik}\} \cdot M\{b_{kj}\} \right\} = M\{A\} \cdot M\{B\},$$

что и требовалось доказать.

*Утверждение 2.* Пусть  $A = \{a_{ij}\}$  и  $B = \{b_{ij}\}$  - случайные матрицы порядка  $(m \times n)$  и  $(n \times r)$  соответственно. Тогда математическое ожидание суммы матриц равно сумме математических ожиданий слагаемых, т.е.  $M\{A+B\} = M\{A\} + M\{B\}$ .

*Доказательство.* На основании определения математического ожидания матрицы заключаем, что

$$M\{A+B\} = \{M\{a_{ij}+b_{ij}\}\} = \{M\{a_{ij}\} + M\{b_{ij}\}\} = M\{A\} + M\{B\},$$

что и требовалось доказать.

Найдем математическое ожидание и ковариационную матрицу вектора  $b^*$  с помощью утверждений 1, 2 и выражения для  $b^{*R}$ , приведенного выше. Имеем

$$M\{b^{*R}\} = b + M\{(X_R^T X_R)^{-1} X_R^T e\} = b + M\{(X_R^T X_R)^{-1} X_R^T\} \cdot M\{e\}.$$

Но так как  $M\{e\} = 0$ , то  $M\{b^{*R}\} = b$ . Это означает что оценка МНК является несмещенной.

Найдем ковариационную матрицу:

$$D\{b^{*R}\} = M\{(b^{*R} - b)(b^{*R} - b)^T\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot e \cdot e^T \cdot X_R (X_R^T X_R)^{-1}\}.$$

Можно доказать, что

$$D\{b^{*R}\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot M\{e \cdot e^T\} \cdot X_R (X_R^T X_R)^{-1}\},$$

но

$$M\{e \cdot e^T\} = D\{e\} = S^2 E,$$

поэтому

$$D\{\hat{b}_R\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot (S^2 E) \cdot X_R (X_R^T X_R)^{-1}\} = S^2 \cdot M\{(X_R^T X_R)^{-1}\}.$$

Как выяснено ранее, для достаточно большого количества опытов  $n$  выполняется приближенное равенство

$$(X_R^T X_R)^{-1} \approx \frac{1}{n} E, \quad (51)$$

тогда

$$D\{b^{*R}\} = \frac{S^2}{n} E.$$

Осталось определить вид распределения вектора  $b^{*R}$ . Из выражения для  $b^{*R}$ , приведенного выше, и асимптотического соотношения (51) следует, что

$$b^{*R} = b + \frac{1}{n} X_R^T e.$$

Можно утверждать, что вектор  $b^{*R}$  имеет асимптотически нормальное распределение, т.е.

$$b^{*R} \in N(b, \frac{S^2}{n} E).$$

Тогда совместная функция плотности распределения вероятностей случайных величин  $b^{*R}_1, b^{*R}_2, \dots, b^{*R}_m$  будет иметь вид:

$$f(b^{*R}) = \frac{1}{(2p)^{m/2} \cdot (\det C)^{1/2}} \cdot \exp[-\frac{1}{2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)], \quad (52)$$

где

$$C = D(b^{*R}) = \frac{S^2}{n} E.$$

Тогда справедливы соотношения

$$C^{-1} = \frac{n}{S^2} E, \quad \det C = \det(\frac{n}{S^2} E) = (\frac{S^2}{n})^m.$$

Подставим в формулу (52), получим

$$\begin{aligned} f(b^{*R}) &= \frac{1}{(2p)^{m/2} \cdot (S^2/n)^{m/2}} \cdot \exp[-\frac{n}{2S^2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)] = \\ &= \frac{1}{(S\sqrt{2p/n})^m} \exp[-\frac{n}{2S^2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)] = \\ &= \frac{1}{(S\sqrt{2p/n})^m} \exp[-\frac{n}{2S^2}(b_1^2 + b_2^2 + \dots + b_m^2)], \end{aligned}$$

где

$$b_i = b_i^{*R} - b_i, \quad i = 1, 2, \dots, m.$$

Вычислим асимптотическую вероятность попадания описывающего реальность вектора параметров  $b$  в  $m$ -мерный куб с длиной стороны, равной  $2k$ , и с центром  $b^{*R}$ .

$$\begin{aligned} P(-k < b_1 < k, -k < b_2 < k, \dots, -k < b_m < k) &= \\ &= \frac{1}{(s\sqrt{2p/n})^m} \left\{ \int_{-k}^k \dots \int_{-k}^k \exp\left[-\frac{n}{2s^2}(b_1^2 + b_2^2 + \dots + b_m^2)\right] \cdot db_1 \dots db_m \right\} = \\ &= \frac{1}{(s\sqrt{2p/n})^m} \left\{ \int_{-k}^k \exp\left[-\frac{n}{2s^2}b_1^2\right] db_1 \dots \int_{-k}^k \exp\left[-\frac{n}{2s^2}b_i^2\right] db_i \right\}. \end{aligned}$$

Сделаем замену

$$t_i = \sqrt{n/2} \cdot \frac{1}{s} b_i, \quad i = 1, 2, \dots, m.$$

Тогда

$$\begin{aligned} P &= P(-k < b_1 < k, -k < b_2 < k, \dots, -k < b_m < k, ) = \\ &= \frac{(s\sqrt{2/n})^m}{(s\sqrt{2p/n})^m} \left[ \int_{-T}^T e^{-t^2} dt \right]^m = \left[ (1/\sqrt{p}) \int_{-T}^T e^{-t^2} dt \right]^m = [\Phi_0(T)]^m, \end{aligned}$$

где  $T = (n/2)^{1/2}(k/\sigma)$ , а  $\Phi_0(T)$ - интеграл Лапласа,

$$\Phi_0(T) = \Phi(\sqrt{2}T) - \Phi(-\sqrt{2}T),$$

где  $\Phi(t)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из последнего соотношения получаем

$$T = \Phi^{-1}(P^{1/m}),$$

где  $\Phi^{-1}(P)$  - обратная функция Лапласа. Отсюда следует, что

$$k = \sigma (2/n)^{1/2} \Phi^{-1}(P^{1/m}). \quad (53)$$

Напомним, что доверительная область  $B_\alpha(n, b)$  - это  $m$ -мерный куб, длина стороны которого равна  $K$ , т.е.

$$P(b \in B_\alpha(n, b)) = P(-K < \beta_1 < K, -K < \beta_2 < K, \dots, -K < \beta_m < K) = \alpha.$$

Подставляя  $P = \alpha$  в формулу (53), получим

$$K = k = \sigma (2/n)^{1/2} \Phi^{-1}(\alpha^{1/m}). \quad (54)$$

Соотношение (54) выражает зависимость размеров доверительной области (т.е. длины ребра куба  $K$ ) от числа опытов  $n$ , среднего квадратического отклонения  $\sigma$  ошибки  $e$  и доверительной вероятности  $\alpha$ . Это соотношение понадобится для определения рационального объема выборки.

Переходим к этапу 2. Необходимо ввести меру разброса (неопределенности) и установить соотношение между статистической и интервальной (метрологической) неопределенностями с соответствии с ранее сформулированным общим подходом.

Пусть  $A$  - некоторое измеримое множество точек в  $m$ -мерном евклидовом пространстве, характеризующее неопределенность задания вектора  $a \hat{I} A$ . Тогда необходимо ввести некую меру  $M(A)$ , измеряющую степень неопределенности. Такой мерой может служить  $m$ -мерный объем  $V(A)$  множества  $A$  (т.е. его мера Лебега или Жордана),  $M(A) = V(A)$ .

Пусть  $P$  -  $m$ -мерный параллелепипед, характеризующий интервальную неопределенность. Длины его сторон равны значениям *нотн*  $2N_1, 2N_2, \dots, 2N_m$ , а центр  $a$  (точка пересечений диагоналей параллелепипеда) находится в точке  $b^{*R}$ . Пусть  $C$  - измеримое множество точек, характеризующее общую неопределенность. В рассматриваемом случае это  $m$ -мерный параллелепипед, длины сторон которого равны  $2(N_1 + K), 2(N_2 + K), \dots, 2(N_m + K)$ , а центр находится в точке  $b^{*R}$ . Тогда

$$M(P) = V(P) = 2^m N_1 N_2 \dots N_m, \quad (55)$$

$$M(C) = V(C) = 2^m (N_1 + K)(N_2 + K) \dots (N_m + K). \quad (56)$$

Справедливо соотношение (49), согласно которому  $M(C) = M(P) + M(F)$ , где множество  $F = C \setminus P$  характеризует статистическую неопределенность.

На этапе 3 получаем по результатам этапов 1 и 2 выражение для рационального объема выборки. Найдем то число опытов, при

котором статистическая неопределенность составит  $\delta$  100% от общей неопределенности, т.е. согласно правилу (50)

$$M(F) = M(C) - M(P) = \delta M(C) \quad (57)$$

где  $0 < d < 1$ . Подставив (55) и (56) в (57), получим

$$2^m \prod_{i=1}^m (N_i + K) - 2^m \prod_{i=1}^m (N_i) = 2^m d \prod_{i=1}^m (N_i + K).$$

Следовательно,

$$(1 - d) \prod_{i=1}^m (N_i + K) / \prod_{i=1}^m (N_i) = 1.$$

Преобразуем эту формулу:

$$(1 - d) \prod_{i=1}^m (1 + K / N_i) = 1,$$

откуда

$$\prod_i^m (1 + K / N_i) = 1 / (1 - d).$$

Если статистическая погрешность мала относительно метрологической, т.е. величины  $K/N_i$  малы, то

$$\prod_i^m (1 + K / N_i) \approx 1 + \sum_i^m (K / N_i).$$

При  $m = 1$  эта формула является точной. Из нее следует, что для дальнейших расчетов можно использовать соотношение

$$1 + \sum_i^m (K / N_i) = 1 / (1 - d).$$

Отсюда нетрудно найти  $K$ :

$$K = \frac{d}{1 - d} \left( 1 / \sum_{i=1}^m (1 / N_i) \right). \quad (58)$$

Подставив в формулу (58) зависимость  $K = K(n)$ , полученную в формуле (54), находим приближенное (асимптотическое) выражение для рационального объема выборки:

$$n_{\text{рац}} = 2 \left( \frac{1-d}{d} s \sum_{i=1}^m (1/N_i) \cdot \Phi^{-1}(a^{1/m}) \right)^2.$$

При  $m = 1$  эта формула также справедлива, более того, является точной.

Переход от произведения к сумме является обоснованным при достаточно малом  $d$ , т.е. при достаточно малой статистической неопределенности по сравнению с метрологической. В общем случае можно находить  $K$  и затем рациональный объем выборки тем или иным численным методом.

**Пример 1.** Представляет интерес определение  $n_{\text{рац}}$  для случая, когда  $m = 2$ , поскольку простейшая линейная регрессия с  $m = 2$  широко применяется. В этом случае базовое соотношение имеет вид

$$(1 + K/N_1)(1 + K/N_2) = 1/(1 - d).$$

Решая это уравнение относительно  $K$ , получаем

$$K = 0.5 \{ -(N_1 + N_2) + [(N_1 + N_2)^2 + 4 N_1 N_2 (d/(1 - d))]^{1/2} \}.$$

Далее, подставив в формулу (54), получим уравнение для рационального объема выборки в случае  $m = 2$ :

$$\sigma (2/n)^{1/2} \Phi^{-1}(a^{1/2}) = 0.5 \{ -(N_1 + N_2) + [(N_1 + N_2)^2 + 4 N_1 N_2 (d/(1 - d))]^{1/2} \}.$$

Следовательно,

$$n_{\text{рат}} = \frac{8 \{ \Phi^{-1}(\sqrt{a}) \}^2}{\left\{ -\frac{N_1}{s} - \frac{N_2}{s} + \sqrt{\left( \frac{N_1}{s} + \frac{N_2}{s} \right)^2 + 4 \frac{N_1 N_2 d}{s^2 (1-d)}} \right\}^2}.$$

При использовании «принципа уравнивания погрешностей» согласно [3]  $d = 1/2$ . При доверительной вероятности  $a = 0,95$  имеем  $\sqrt{a} = 0,9747$  и согласно [42]  $\Phi^{-1}(\sqrt{a}) = 1,954$ . Для этих численных значений

$$n_{rat} = \frac{30,545}{\left\{ -\frac{N_1}{S} - \frac{N_2}{S} + \sqrt{\left(\frac{N_1}{S} + \frac{N_2}{S}\right)^2 + 4\frac{N_1 N_2}{S^2}} \right\}^2}.$$

Если  $\frac{N_1}{S} = \frac{N_2}{S} = 0,1$ , то  $n_{rat} = 4455$ . Если же  $\frac{N_1}{S} = \frac{N_2}{S} = 0,5$ , то  $n_{rat} = 178$ .

Если первое из этих чисел превышает обычно используемые объемы выборок, то второе находится в «рабочей зоне» регрессионного анализа.

**Парная регрессия.** Наиболее простой и одновременно наиболее широко применяемый частный случай парной регрессии рассмотрим подробнее. Модель имеет вид

$$y_i = ax_i + b + e_i, \quad i = 1, 2, \dots, n.$$

Здесь  $x_i$  – значения фактора (независимой переменной),  $y_i$  – значения отклика (зависимой переменной),  $e_i$  – статистические погрешности,  $a, b$  – неизвестные параметры, оцениваемые методом наименьших квадратов. Она переходит в модель (используем альтернативную запись линейной модели)

$$y = Xb + e,$$

если положить

$$y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 & 1 \\ \dots & \dots \\ x_n & 1 \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ \dots \\ e_n \end{pmatrix}, \quad b = \begin{pmatrix} a \\ b \end{pmatrix}$$

Естественно принять, что погрешности факторов описываются матрицей

$$\Delta X = a = \begin{pmatrix} \Delta x_1 & 0 \\ \dots & \dots \\ \Delta x_n & 0 \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ \dots & \dots \\ a_n & 0 \end{pmatrix}$$

В рассматриваемой модели интервального метода наименьших квадратов

$$X = X_R + a, \quad y = y_R + g \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$$

где  $X, y$  – наблюдаемые (т.е. известные статистику) значения фактора и отклика,  $X_R, y_R$  – истинные значения переменных,  $a, g$  – погрешности измерений переменных. Пусть  $b^*$  – оценка метода наименьших квадратов, вычисленная по наблюдаемым значениям переменных,  $b_R^*$  – аналогичная оценка, найденная по истинным значениям. В соответствии с ранее проведенными рассуждениями

$$b^* - b = [-(X_0^T X_0)^{-1} \Delta (X_0^T X_0)^{-1} X_0^T + (X_0^T X_0)^{-1} a^T] y_0 + (X_0^T X_0)^{-1} X_0^T g \quad (59)$$

с точностью до бесконечно малых более высокого порядка по  $|a|$  и  $|g|$ . В формуле (59) использовано обозначение  $\Delta = X_0^T a + a^T X_0$ . Вычислим правую часть в (59), выделим главный линейный член и найдем нотну.

Легко видеть, что

$$X^T X = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix} \quad (60)$$

где суммирование проводится от 1 до  $n$ . Для упрощения обозначений в дальнейшем до конца настоящего пункта не будем указывать эти пределы суммирования. Из (60) вытекает, что

$$(X^T X)^{-1} = \begin{pmatrix} n & -\sum x_i \\ -\sum x_i & \sum x_i^2 \end{pmatrix} / [n \sum x_i^2 - (\sum x_i)^2]. \quad (61)$$

Легко подсчитать, что

$$X^T a + a^T X = \begin{pmatrix} 2 \sum x_i a_i & \sum a_i \\ \sum a_i & n \end{pmatrix} \quad (62)$$

Положим

$$S_0^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

Тогда знаменатель в (61) равен  $n^2 S_0^2$ . Из (61) и (62) следует, что

$$(X^T X)^{-1} (X^T a + a^T X) = \frac{1}{n^2 S_0^2} \begin{pmatrix} 2n \sum x a - \sum x \sum a & n \sum a \\ -2 \sum x \sum x a + \sum x^2 \sum a & -\sum x \sum a \end{pmatrix} \quad (63)$$

Здесь и далее опустим индекс  $i$ , по которому проводится суммирование. Это не может привести к недоразумению, поскольку всюду суммирование проводится по индексу  $i$  в интервале от 1 до  $n$ . Из (61) и (63) следует, что

$$(X^T X)^{-1}(X^T a + a^T X)(X^T X)^{-1} = \frac{1}{n^4 S_0^4} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad (64)$$

где

$$\begin{aligned} A &= 2n^2 \sum xa - 2n \sum x \sum a, \\ B = C &= -2n \sum x \sum xa + (\sum x)^2 \sum a + n \sum a \sum x^2, \\ D &= 2(\sum x)^2 \sum xa - 2 \sum a \sum x \sum x^2. \end{aligned}$$

Наконец, вычисляем основной множитель в (59)

$$(X^T X)^{-1}(X^T a + a^T X)(X^T X)^{-1} X^T = \frac{1}{n^4 S_0^4} \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1i} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2i} & \dots & z_{2n} \end{pmatrix} \quad (65)$$

где

$$z_{1i} = Ax_i + B, \quad z_{2i} = Cx_i + D, \quad i = 1, 2, \dots, n.$$

Перейдем к вычислению второго члена с  $a$  в (59). Имеем

$$(X^T X)^{-1} a^T = \frac{1}{n^2 S_0^2} \begin{pmatrix} w_{11} & \dots & w_{1i} & \dots & w_{1n} \\ w_{21} & \dots & w_{2i} & \dots & w_{2n} \end{pmatrix} \quad (67)$$

где

$$w_{1i} = na_i, \quad w_{2i} = -a_i \sum x, \quad i = 1, 2, \dots, n.$$

Складывая правые части (65) и (67) и умножая на  $y$ , получим окончательный вид члена с  $a$  в (59):

$$\{(X^T X)^{-1}(X^T a + a^T X)(X^T X)^{-1} X^T + (X^T X)^{-1} a^T\} y = \begin{pmatrix} F \\ G \end{pmatrix} \quad (68)$$

где

$$\begin{aligned} F &= (\sum xy) (2n^2 \sum xa - 2n \sum x \sum a) / n^4 S_0^4 + (\sum ya) / n S_0^2 + \\ &\quad + (\sum y) (n \sum a \sum x^2 + \sum a (\sum x)^2 - 2n \sum x \sum xa) / n^4 S_0^4, \\ G &= (\sum xy) (-2n \sum x \sum xa + n \sum a \sum x^2 + \sum a (\sum x)^2) / n^4 S_0^4 - \\ &\quad - (\sum ya) (\sum x) / n^2 S_0^2 + (\sum y) (2 \sum xa (\sum x)^2 - 2 \sum a \sum x \sum x^2) / n^4 S_0^4. \end{aligned} \quad (69)$$

Для вычисления нотны выделим главный линейный член.

Сначала найдем частные производные. Имеем

$$\begin{aligned}\frac{\partial F}{\partial a_j} &= (\sum xy)(2n^2 x_j - 2n \sum x) / n^4 S_0^4 + y_j / n S_0^2 + \\ &+ (\sum y)(n \sum x^2 + (\sum x)^2 - 2n(\sum x)x_j) / n^4 S_0^4; \\ \frac{\partial G}{\partial a_j} &= (\sum xy)(-2n(\sum x)x_j + n \sum x^2 + (\sum x)^2) / n^4 S_0^4 - \\ &- y_j (\sum x) / n^2 S_0^2 + (\sum y)(2x_j (\sum x)^2 - 2 \sum x \sum x^2) / n^4 S_0^4.\end{aligned}\quad (70)$$

Если ограничения имеют вид

$$|a_j| \leq \Delta, \quad j = 1, 2, \dots, n,$$

то максимально возможное отклонение оценки  $a^*$  параметра  $a$  из-за погрешностей  $a_j$  таково:

$$N_a(x) = \sum_{1 \leq j \leq n} \left| \frac{\partial F}{\partial a_j} \right| \Delta + O(\Delta^2),$$

где производные заданы формулой (70).

**Пример 2.** Пусть вектор  $(x, y)$  имеет двумерное нормальное распределение с нулевыми математическими ожиданиями, единичными дисперсиями и коэффициентом корреляции  $r$ . Тогда

$$\lim_{\Delta \rightarrow 0} \lim_{n \leftarrow \infty} \frac{N_a(x)}{\Delta} = \lim_{n \rightarrow \infty} \sum_{1 \leq j \leq n} \left| \frac{\partial F}{\partial a_j} \right| = M |2rx + y| = \sqrt{\frac{2(1+8r^2)}{p}}. \quad (71)$$

При этом

$$\lim_{n \rightarrow \infty} \frac{\partial G}{\partial a_j} = r,$$

следовательно, максимально возможному изменению параметра  $b^*$  соответствует сдвиг всех  $x_i$  в одну сторону, т.е. наличие систематической ошибки при определении  $x$ -ов. В то же время согласно (71) значения  $a_j$  в асимптотике выбираются по правилу

$$a_j = \begin{cases} \Delta, & 2rx_j + y_j > 0, \\ -\Delta, & 2rx_j + y_j \leq 0. \end{cases}$$

Таким образом, максимальному изменению  $a^*$  соответствуют не те  $a_j$ , что максимальному изменению  $b^*$ . В этом – новое по сравнению с одномерным случаем. В зависимости от вида ограничений на возможные отклонения, в частности, от вида метрики в пространстве параметров, будут «согласовываться» отклонения по отдельным параметрам. Ситуация аналогична той, что возникает в классической математической статистике в связи с оптимальным оцениванием параметров. Если параметр одномерен, то ситуация с оцениванием достаточно прозрачна – есть понятие эффективных оценок, показателем качества оценки является средний квадрат ошибки, а при ее несмещенности – дисперсия. В случае нескольких параметров возникает необходимость соизмерить точность оценивания по разным параметрам. Есть много критериев оптимальности (см., например, [46]), но нет признанных правил выбора среди них.

Вернемся к формуле (59). Интересно, что отклонения вектора параметров, вызванные отклонениями значений факторов  $a$  и отклика  $g$ , входят в (59) аддитивно. Хотя

$$\sup_{a,g} \| b^* - b \| \neq \sup_a | \{ -(X_0^T X_0)^{-1} \Delta (X_0^T X_0)^{-1} X_0^T + (X_0^T X_0)^{-1} a^T \} y_0 | + \\ + \sup_g | (X_0^T X_0)^{-1} X_0^T g |,$$

но для отдельных компонент (не векторов!) имеет место равенство.

В случае парной регрессии

$$(X_0^T X_0)^{-1} X_0^T g = \frac{1}{n^2 S_0^2} \left( \sum g_i (n x_i - \sum x); \sum g_i (-x_i \sum x + \sum x^2) \right)^T. \quad (72)$$

Из формул (68), (69) и (72) следует, что

$$b^* - b = \begin{pmatrix} a^*(X, y) - a^*(X_0, y_0) \\ b^*(X, y) - b^*(X_0, y_0) \end{pmatrix} = \begin{pmatrix} F + F_1 \\ G + G_1 \end{pmatrix}$$

где  $F$  и  $G$  определены в (69), а

$$F_1 = \frac{1}{n^2 S_0^2} \left( n \sum g^x - \sum x \sum g \right), \quad G_1 = \frac{1}{n^2 S_0^2} \left( \sum g \sum x^2 - \sum g^x \sum x \right).$$

Итак, продемонстрирована возможность применения основных подходов статистики интервальных данных в регрессионном анализе.

### 2.3.8. Интервальный дискриминантный анализ

Перейдем к задачам классификации в статистике интервальных данных. Как известно [27], важная их часть – задачи дискриминации (диагностики, распознавания образов с учителем). В этих задачах заданы классы (полностью или частично, с помощью обучающих выборок), и необходимо принять решение – к какому этих классов отнести вновь поступающий объект.

В линейном дискриминантном анализе правило принятия решений основано на линейной функции  $f(x)$  от распознаваемого вектора  $x \in R^k$ . Рассмотрим для простоты случай двух классов. Правило принятия решений определяется константой  $C$  – при  $f(x) > C$  распознаваемый объект относится к первому классу, при  $f(x) \leq C$  – ко второму.

В первоначальной вероятностной модели Р.Фишера предполагается, что классы заданы обучающими выборками объемов  $N_1$  и  $N_2$  соответственно из многомерных нормальных распределений с разными математическими ожиданиями, но одинаковыми ковариационными матрицами. В соответствии с леммой Неймана-Пирсона, дающей правило принятия решений при проверке статистических гипотез, дискриминантная функция является линейной. Для ее практического использования теоретические характеристики распределения необходимо заменить на выборочные. Тогда дискриминантная функция приобретает следующий вид

$$f(x) = \left( x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2) \right)^T S^{-1}(\bar{x}_1 - \bar{x}_2).$$

Здесь  $\bar{x}_1$  - выборочное среднее арифметическое по первой выборке  $x_a^{(1)}$ ,  $a = 1, 2, \dots, N_1$ , а  $\bar{x}_2$  - выборочное среднее арифметическое по второй выборке  $x_b^{(2)}$ ,  $b = 1, 2, \dots, N_2$ . В роли  $S$  может выступать любая состоятельная оценка общей для выборок ковариационной матрицы. Обычно используют следующую оценку, естественным образом сконструированную на основе выборочных ковариационных матриц:

$$S = \frac{\sum_{a=1}^{N_1} (x_a^{(1)} - \bar{x}_1)(x_a^{(1)} - \bar{x}_1)^T + \sum_{b=1}^{N_2} (x_b^{(2)} - \bar{x}_2)(x_b^{(2)} - \bar{x}_2)^T}{N_1 + N_2 - 2}.$$

В соответствии с подходом статистики интервальных данных считаем, что специалисту по анализу данных известны лишь значения с погрешностями

$$y_a^{(1)} = x_a^{(1)} + e_a^{(1)}, \quad a = 1, 2, \dots, N_1, \quad y_b^{(2)} = x_b^{(2)} + e_b^{(2)}, \quad b = 1, 2, \dots, N_2.$$

Таким образом, вместо  $f(x)$  статистик делает выводы на основе искаженной линейной дискриминантной функции  $f_l(x)$ , в которой коэффициенты рассчитаны не по исходным данным  $x_a^{(1)}, x_b^{(2)}$ , а по искаженным погрешностями значениям  $y_a^{(1)}, y_b^{(2)}$ .

Это – модель с искаженными параметрами дискриминантной функции. Следующая модель – такая, в которой распознаваемый вектор  $x$  также известен с ошибкой. Далее, константа  $C$  может появляться в модели различными способами. Она может задаваться априори абсолютно точно. Может задаваться с какой-то ошибкой, не связанной с ошибками, вызванными конечностью обучающих выборок. Может рассчитываться по обучающим выборкам, например, с целью уравнивать ошибки классификации, т.е. провести плоскость дискриминации через середину отрезка, соединяющего центры классов. Итак – целый спектр моделей ошибок.

На какие статистические процедуры влияют ошибки в исходных данных? Здесь тоже много постановок. Можно изучать

влияние погрешностей измерений на значения дискриминантной функции  $f$ , например, в той точке, куда попадает вновь поступающий объект  $x$ . Очевидно, случайная величина  $f(x)$  имеет некоторое распределение, определяемое распределениями обучающих выборок. Выше описана модель Р.Фишера с нормально распределенными совокупностями. Однако реальные данные, как правило, не подчиняются нормальному распределению [27]. Тем не менее линейный статистический анализ имеет смысл и для распределений, не являющихся нормальными (при этом вместо свойств многомерного нормального распределения приходится опираться на многомерную центральную предельную теорему и теорему о наследовании сходимости [3]). В частности, приравняв метрологическую ошибку, вызванную погрешностями исходных данных, и статистическую ошибку, получим условие, определяющее рациональность объемов выборок. Здесь два объема выборок, а не один, как в большинстве рассмотренных постановок статистики интервальных данных. С подобным мы сталкивались ранее при рассмотрении двухвыборочного критерия Смирнова.

Естественно изучать влияние погрешностей исходных данных не при конкретном  $x$ , а для правила принятия решений в целом. Может представлять интерес изучение характеристик этого правила по всем  $x$  или по какому-либо отрезку. Более интересно рассмотреть показатель качества классификации, связанный с пересчетом на модель линейного дискриминантного анализа [27].

Математический аппарат изучения перечисленных моделей развит выше в предыдущих пунктах настоящей главы. Некоторые результаты приведены в [14]. Из-за большого объема выкладок ограничимся приведенными здесь замечаниями.

### **2.3.9. Интервальный кластер-анализ**

Кластер-анализ, как известно [27], имеет целью разбиение совокупности объектов на группы сходных между собой. Многие методы кластер-анализа основаны на использовании расстояний между объектами. (Степень близости между объектами может измеряться также с помощью мер близости и показателей различия, для которых неравенство треугольника выполнено не всегда.) Рассмотрим влияние погрешностей измерения на расстояния между объектами и на результаты работы алгоритмов кластер-анализа.

С ростом размерности  $p$  евклидова пространства диагональ единичного куба растет как  $\sqrt{p}$ . А какова погрешность определения евклидова расстояния? Пусть двум рассматриваемым векторам соответствуют  $X_0 = (x_1, x_2, \dots, x_p)$  и  $Y_0 = (y_1, y_2, \dots, y_p)$  - вектора размерности  $p$ . Они известны с погрешностями  $e = (e_1, e_2, \dots, e_p)$  и  $d = (d_1, d_2, \dots, d_p)$ , т.е. статистику доступны лишь вектора  $X = X_0 + e$ ,  $Y = Y_0 + d$ . Легко видеть, что

$$r^2(X, Y) = r^2(X_0, Y_0) + 2 \sum_{1 \leq i \leq p} (x_i - y_i)(e_i - d_i) + \sum_{1 \leq i \leq p} (e_i - d_i)^2. \quad (73)$$

Пусть ограничения на абсолютные погрешности имеют вид

$$|e_i| \leq \Delta, \quad |d_i| \leq \Delta, \quad i = 1, 2, \dots, n.$$

Такая запись ограничений предполагает, что все переменные имеют примерно одинаковый разброс. Трудно ожидать этого, если переменные имеют различные размерности. Однако рассматриваемые ограничения на погрешности естественны, если переменные предварительно стандартизованы, т.е. отнормированы (т.е. из каждого значения вычтено среднее арифметическое, а разность поделена на выборочное среднее квадратическое отклонение).

Пусть  $p\Delta^2 \rightarrow 0$ . Тогда последнее слагаемое в (73) не превосходит  $4p\Delta^2$ , поэтому им можно пренебречь. Тогда из (73) следует, что нотна евклидова расстояния имеет вид

$$N_{r^2}(X_0, Y_0) = 4 \sum_{1 \leq i \leq p} |x_i - y_i| \Delta$$

с точностью до бесконечно малых более высокого порядка. Если случайные величины  $|x_i - y_i|$  имеют одинаковые математические ожидания и для них справедлив закон больших чисел (эти предположения естественны, если переменные перед применением кластер-анализа стандартизованы), то существует константа  $C$  такая, что

$$N_{r^2}(X_0, Y_0) = Cp\Delta$$

с точностью до бесконечно малых более высокого порядка при малых  $\Delta$ , больших  $p$  и  $p\Delta^2 \rightarrow 0$ .

Из рассмотрений настоящего пункта вытекает, что

$$r(X, Y) = r(X_0, Y_0) + q \frac{Cp\Delta}{2r(X_0, Y_0)} \quad (74)$$

при некотором  $q$  таком, что  $|q| < 1$ .

Какое минимальное расстояние является различимым? По аналогии с определением рационального объема выборки при проверке гипотез предлагается уравнивать слагаемые в (74), т.е. определять минимально различимое расстояние  $r_{\min}$  из условия

$$r_{\min} = \frac{Cp\Delta}{2r_{\min}}, \quad r_{\min} = \sqrt{\frac{Cp\Delta}{2}}. \quad (75)$$

Естественно принять, что расстояния, меньшие  $r_{\min}$ , не отличаются от 0, т.е. точки, лежащие на расстоянии  $r \leq r_{\min}$ , не различаются между собой.

Каков порядок величины  $C$ ? Если  $x_i$  и  $y_i$  независимы и имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то, как легко подсчитать,  $M|x_i - y_i| = 2/\sqrt{p} = 1,13$  и соответственно  $C = 4,51$ . Следовательно, в этой модели

$$r_{\min} = 1,5\sqrt{p\Delta}.$$

Формула (75) показывает, что хотя с ростом размерности пространства  $p$  растет диаметр (длина диагонали) единичного куба – естественной области расположения значений переменных, с той же скоростью растет и естественное квантование расстояния с помощью порога неразличимости  $r_{\min}$ , т.е. увеличение размерности (вовлечение новых переменных), вообще говоря, не улучшает возможности кластер-анализа.

Можно сделать выводы и для конкретных алгоритмов. В дендрограммах (например, результатах работы иерархических агломеративных алгоритмах ближнего соседа, дальнего соседа, средней связи) можно порекомендовать склеивать (т.е. объединять) уровни, отличающиеся менее чем на  $r_{\min}$ . Если все уровни склеятся, то можно сделать вывод, что у данных нет кластерной структуры, они однородны. В алгоритмах типа «Форель» центр тяжести текущего кластера определяется с точностью  $\pm \Delta$  по каждой координате, а порог для включения точки в кластер (радиус шара  $R$ ) из-за погрешностей исходных данных может измениться согласно (74) на

$$\pm \frac{2,25}{R} p\Delta.$$

Поэтому кроме расчетов с  $R$  рекомендуется провести также расчеты с радиусами  $R_1$  и  $R_2$ , где

$$R_1 = R \left( 1 - \frac{2,25}{R^2} p\Delta \right), \quad R_2 = R \left( 1 + \frac{2,25}{R^2} p\Delta \right),$$

и сравнить полученные разбиения. Быть адекватными реальности могут только выводы, общие для всех трех расчетов. Эти рекомендации развивают общую идею [3] о целесообразности проведения расчетов при различных значениях параметров алгоритмов с целью выделения выводов, инвариантных по отношению к выбору конкретного алгоритма.

### **2.3.10. Место статистики интервальных данных (СИД) среди методов описания неопределенностей**

Кратко рассмотрим положение статистики интервальных данных среди других методов описания неопределенностей.

**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 1 в некотором интервале и 0 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 продвинуться дальше, чем при использовании функций принадлежности произвольного вида.

**Интервальная математика и СИД.** Можно было бы сказать, что СИД – часть интервальной математики, что СИД 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 формулы для выборочной дисперсии, рассмотренные выше, имеют разные погрешности. А с точки зрения СИД эти две формулы задают одну и ту же функцию, и 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 дело будущего. Из уже сделанного отметим

применение методов СИД при анализе такой характеристики финансовых потоков, как  $NPV$  – чистая текущая стоимость [27].

**Математическая статистика и СИД.** Как уже отмечалось, математическая статистика и СИД отличаются тем, в каком порядке делаются предельные переходы  $n \rightarrow \infty$  и  $\Delta \rightarrow 0$ . При этом СИД переходит в математическую статистику при  $\Delta = 0$ . Правда, тогда исчезают основные особенности СИД: нотна становится равной 0, а рациональный объем выборки – бесконечности. Рассмотренные выше методы СИД разработаны в предположении, что погрешности малы (но не исчезают) и 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) расширяет область применения классической математической статистики путем включения в нее новых видов статистических данных [27]. 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 классике, чем другие части СОНП, например, статистика бинарных отношений.

**Робастные методы статистики и СИД.** Если понимать робастность согласно [3] как теорию устойчивости статистических

методов по отношению к допустимым отклонениям исходных данных и предпосылок модели, 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 [27]. К этой модели СИД не имеет отношения.

**Теория устойчивости и СИД.** Общей схеме устойчивости [3] математических моделей социально-экономических явлений и процессов по отношению к допустимым отклонениям исходных данных и предпосылок моделей СИД полностью соответствует. Он посвящен 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 устойчивости [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, используемый, 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], 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 (см. выше). Подчеркнем, что применение в СИД вероятностно-статистических моделей погрешностей не менее перспективно, чем минимаксных.

**Подход научной школы А.П. Воцинина и СИД.** Если в математической статистике неопределенность только статистическая, то в научной школе А.П. Воцинина - только интервальная. Можно сказать, что СИД лежит между классической прикладной математической статистикой и областью исследований научной школы А.П. Воцинина. Другое отличие состоит в том, что в этой школе разрабатывают новые методы анализа интервальных данных, а в СИД в настоящее время изучается 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  $\pm 10\%$  от базового значения. Однако этот анализ делают по каждой переменной отдельно. В СИД все переменные рассматриваются совместно, и находится максимально возможное отклонение (нотна). 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.

## Литература

1. Дискуссия по анализу интервальных данных // Заводская лаборатория. 1990. Т.56. No.7, с.75-95.
2. Сборник трудов Международной конференции по интервальным и стохастическим методам в науке и технике (ИНТЕРВАЛ-92). Тт. 1,2. - М.: МЭИ, 1992, 216 с., 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 Computations, 1992, No.1(3), p.44-52.
6. Орлов А.И. // Наука и технология в России. 1994. No.4(6). С.8-9.
7. Шокин Ю.И. Интервальный анализ. Новосибирск: Наука, 1981, 112 с.
8. Орлов А.И. - 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. - 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. - 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. - 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. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. Пермь: Пермский государственный университет, 1993, с.149-158.
13. Битгар А.Б. Метод наименьших квадратов для интервальных данных. Дипломная работа. - М.: МЭИ, 1994. 38 с.

14. Пузикова Д.А. // Наука и технология в России. 1995. №2(8). С.12-13.
15. Орлов А.И. // Надежность и контроль качества, 1991, № 8, с.3-8.
16. Орлов А.И. // Заводская лаборатория. 1998. Т.64. № 3. С.52-60.
17. Вошинин А.П. Метод оптимизации объектов по интервальным моделям целевой функции. - М.: МЭИ, 1987. 109 с.
18. Вошинин А.П., Сотиров Г.Р. Оптимизация в условиях неопределенности. - М.: МЭИ - София: Техника, 1989. 224 с.
19. Вошинин А.П., Акматбеков Р.А. Оптимизация по регрессионным моделям и планирование эксперимента. - Бишкек: Илим, 1991. 164 с.
20. Вошинин А.П. // Заводская лаборатория. 2000. Т.66, № 3. С.51-65.
21. Вошинин А.П. // Заводская лаборатория. 2002. Т.68, № 1. С.118-126.
22. Дывак Н.П. Разработка методов оптимального планирования эксперимента и анализа интервальных данных. Автореф. дисс. канд. технич. наук. - М.: МЭИ, 1992. 20 с.
23. 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.
24. Орлов А.И. // Заводская лаборатория. 1999. Т.65. № 7. С.46-54.
25. Орлов А.И. // 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. Орлов А.И. Эконометрика. – М.: Экзамен, 2002. 576 с.
28. Дейвид Г. Порядковые статистики. – М.: Наука, 1979.
29. Колмогоров А.Н. Метод медианы в теории ошибок. – В кн.: Колмогоров А.Н. Теория вероятностей и математическая статистика: [Сб. статей]. – М.: Наука, 1986. – С.111-114.

30. Орлов А.И. Об оценивании параметров гамма-распределения. - Журнал "Обозрение прикладной и промышленной математики". 1997. Т.4. Вып.3. С.471-482.
31. Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. – М.: Наука, 1970.
32. Бронштейн И.Н., Семендяев К.А. Справочник по математике для инженеров и учащихся втузов. – М.-Л.: ГИТТЛ, 1945.
33. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973. 900 с.
34. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. – М.: ВНИИС, 1987.
35. Ляшенко Н.Н., Никулин М.С. Машинное умножение и деление независимых случайных величин // Записки научных семинаров Ленингр. Отделения Математического ин-та АН СССР, 1986, Т.153.
36. Хьюбер П. Робастность в статистике. – М.: Мир, 1984. 303 с.
37. Орлов А.И. Асимптотика решений экстремальных статистических задач // Анализ нечисловых данных в системных исследованиях. Сб. трудов. Вып.10. – М.: ВНИИ системных исследований АН СССР, 1982. – С.4-12.
38. Крамер Г. Математические методы статистики. – М.: Мир, 1975. 648 с.
39. Боровков А.А. Математическая статистика. – М.: Наука, 1984. 472 с.
40. Кендалл М., Стьюарт А. Теория распределений. – М.: Наука, 1966.
41. Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // Бюллетень МГУ. Сер.А. 1939. Т.2. №2.
42. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.

43. Орлов А.И. О критериях Колмогорова и Смирнова // Заводская лаборатория. 1995. Т.61. No.7. С.59-61.
44. Гантмахер Ф.Р. Теория матриц. - М.: Наука, 1966. -576 с.
45. Розанов Ю.А. Теория вероятностей, случайные процессы и математическая статистика. - М.: Наука, 1989. - 320 с.
46. Налимов В.В., Голикова Т.И. Логические основания планирования эксперимента. – М.: Металлургия, 1976. 128 с.

### **Контрольные вопросы и задачи**

1. Покажите на примерах, что в задачах принятия решений исходные данные часто имеют интервальный характер.
2. В чем особенности подхода статистики интервальных данных в задачах оценивания параметров?
3. В чем особенности подхода статистики интервальных данных в задачах проверки гипотез?
4. Какие новые нюансы проявляются в статистике интервальных данных при переходе к многомерным задачам?

5. Выполните операции над интервальными числами:

вариант 1 - а)[1,2]+[3,4], б)[4,5]-[2,3], в)[3,4]x[5,7], г)[10,20]:[4,5];

вариант 2 - а)[0,2]+[3,5], б)[3,5]-[2,4], в)[2,4]x[5,8], г)[15,25]:[1,5].

6. Выпишите формулу для асимптотической нотны (ошибки по абсолютной величине не превосходят константы  $t$ , предполагающейся малой) для функции

$$f(x_1, x_2) = 5 (x_1)^2 + 10 (x_2)^2 + 7 x_1 x_2.$$

Вычислите асимптотическую нотну в точке  $(x_1, x_2) = (1, 2)$  при  $t = 0, 1$ .

7. Выпишите формулу для асимптотической нотны (ошибки по абсолютной величине не превосходят константы  $t$ , предполагающейся малой) для функции

$$f(x_1, x_2) = 4 (x_1)^2 + 12 (x_2)^2 - 3 x_1 x_2.$$

Вычислите асимптотическую нотну в точке  $(x_1, x_2) = (2, 1)$  при  $t = 0,05$ .

### **Темы докладов, рефератов, исследовательских работ**

1. Классическая математическая статистика как предельный случай статистики интервальных данных.
2. Концепция рационального объема выборки.
3. Сравнение методов оценивания параметров и характеристик распределений в статистике интервальных данных и в классической математической статистике.
4. Подход к проверке гипотез в статистике интервальных данных.
5. Метод наименьших квадратов для интервальных данных.
6. Различные способы учета погрешностей исходных данных в статистических процедурах.
7. Статистика интервальных данных как часть теории устойчивости (с использованием монографии [3]).

