Б. П. ДЕМИДОВИЧ и И. А. МАРОН

ОСНОВЫ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

ИЗДАНИЕ ЧЕТВЕРТОЕ,
ИСПРАВЛЕННОЕ

Допущено Министерством
высшего и среднего специального образования СССР
в качестве учебного пособия
для студентов высших технических учебных заведений

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1970

518
Д 30
УДК 518 0

АННОТАЦИЯ

Книга посвящена изложению важнейших методов и приемов вычислительной математики на базе общего втузовского курса высшей математики. Основная часть книги является учебным пособием по курсу приближенных вычислений для втузов. Книга может быть полезна также для лиц, работающих в области прикладной математики.

Борис Павлович Демидович и Исаак Абрамович Марон

Основы вычислительной математики

М., 1970 г., 664 стр. с илл.

Редактор *А. З. Рывкин*

Техн. редактор *К. Ф. Брудно* Корректор *И. Я. Кришталь*

Печать с матриц. Подписано к печати 19/XII 1969 г. Бумага 60×90¹/₁₆. Физ. печ. л. 41,5. Услови. печ. л. 41,5. Уч.-изд. л. 41,06. Тираж 60 000 экз. Т 15992. Цена книги 1 р. 54 к. Заказ № 939

Издательство «Наука»
Главная редакция физико-математической литературы
Москва, В-71, Ленинский проспект, 15

Ордена Трудового Красного Знамени Ленинградская типография № 1 «Печатный Двор» им. А. М. Горького Главполиграфпрома Комитета по печати при Совете Министров СССР. г. Ленинград, Гатчинская ул., 26.

2-2-4
127-70

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Общие правила вычислительной работы

При выполнении массовых вычислений важно придерживаться определенных простых правил, выработанных практикой, соблюдение которых экономит труд вычислителя и позволяет рационально использовать имеющуюся вычислительную технику и вспомогательные средства.

Прежде всего вычислитель должен разработать подробную *вычислительную схему*, точно указывающую порядок действий и дающую возможность получить искомый результат наиболее простым и быстрым путем. Это особенно необходимо при однотипных вычислениях, так как такая схема, автоматизируя вычисления, позволяет выполнять их более быстро и надежно, что с пользой окупает время, затраченное на составление схемы. Кроме того, имея детальную вычислительную схему для решения задачи, можно использовать труд менее квалифицированных вычислителей.

Составление вычислительной схемы проиллюстрируем на следующем примере. Пусть требуется вычислить значения аналитически заданной функции

$$y = f(x)$$

для заданных значений аргумента $x = x_1, x_2, \ldots, x_n$. Если число этих значений велико, то неразумно вычислять отдельно сначала значение $f(x_1)$, затем значение $f(x_2)$ и т. д., каждый раз выполняя всю совокупность операций, указанных символом f. Гораздо целесообразнее, расчленив функцию f на элементарные операции (*действия*)

$$f(x) = f_m(\ldots(f_2(f_1(x)))\ldots),$$

вычисления производить однотипными операциями:

$$\begin{aligned}
u_i &= f_1(x_i) \qquad (i = 1, 2, \ldots, n);\\
v_i &= f_2(u_i) \qquad (i = 1, 2, \ldots, n);\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
y &= f_m(w_i) \qquad (i = 1, 2, \ldots, n),
\end{aligned}$$

выполняя одну и ту же операцию $f_j (j = 1, 2, \ldots, m)$ для всех рассматриваемых значений аргумента. При этом широко могут быть использованы соответствующие таблицы функций и специализированные счетные машины. Запись результатов вычислений следует производить на особых вычислительных *бланках* или *формулярах*, представляющих собой специальным образом разграфленные и размеченные листы бумаги (применительно к выбранной вычислительной схеме!). На этих бланках, в строго определенных местах, заносятся промежуточные результаты по мере их получения, а также окончательные результаты.

Вычислительные бланки обычно строятся таким образом, чтобы результаты каждой серии однотипных операций заносились в один столбец или в одну строку, причем расположение записей промежуточных результатов должно быть удобным для производства последующих вычислений.

Например, для составления таблицы значений функции

$$y = \frac{e^x + \cos x}{1 + x^2} + \sqrt{1 + \sin^2 x} \qquad (1)$$

можно рекомендовать вычислительный бланк, приведенный в таблице 1.

Таблица 1

Вычислительный бланк для функции (1)

x	$\frac{x^2}{(1)^2}$	e^x	$\sin x$	$\cos x$	$e^x + \cos x$ (3)+(5)	$1 + x^2$ (1)+(2)	$\frac{e^x + \cos x}{1 + x^2}$ (6):(7)	$\sin^2 x$ $(4)^2$	$1 + \sin^2 x$ (1)+(9)	$\frac{\sqrt{1 + \sin^2 x}}{\sqrt{(10)}}$	y (8)++(11)
1	2	3	4	5	6	7	8	9	10	11	12

значения искомой функции y. При правильно составленном бланке вычислитель в процессе вычисления уже фактически не пользуется формулой, по которой ведется расчет, его внимание сосредоточено исключительно на последовательном заполнении столбцов.

Заметим, что расчетная схема и форма бланка существенно зависят от используемой техники вычислений и вспомогательных таблиц. Так, например, в некоторых случаях отдельные промежуточные результаты хранятся в памяти машины и в бланк не заносятся. Иногда стандартные совокупности операций удобно рассматривать как отдельное действие. Например, при использовании логарифмической линейки численное значение выражения вида

$$\frac{ab}{c}$$

можно вычислять сразу, не фиксируя промежуточный результат, и поэтому нет необходимости расчленять его на простейшие операции умножения и деления. Аналогично при работе на электрических счетных машинах процесс отыскания суммы парных произведений

$$\sum_{k=1}^{n} a_k b_k$$

является единым действием. Во многих случаях выгодно преобразовывать данные выражения к специальному искусственному виду (например, заменять деление умножением на обратную величину, или приводить выражение к виду, удобному для логарифмирования, и т. п.).

Второе, на что нужно обратить серьезное внимание, — это *контроль вычислений*. Без проверки вычисление не может считаться законченным. Контроль разделяется на *текущий* и *заключительный*. При текущем контроле, производя добавочные действия, мы с большей или меньшей степенью достоверности убеждаемся, что полученные промежуточные результаты правильны. В противном случае производится пересчет соответствующего этапа. При заключительном контроле проверяется лишь окончательный результат. Например, если вычисляется корень уравнения, то найденное значение можно подставить в уравнение и таким образом узнать, правильно или нет решена задача. По здравому смыслу ясно, что если вычисление очень большое, то рискованно ставить под угрозу всю вычислительную работу, дожидаясь проверки окончательного результата. Поэтому целесообразно проверять правильность расчетов по этапам. В ответственных случаях расчеты контролируются путем независимого выполнения расчетов двумя различными вычислителями, или же задача решается одним и тем же вычислителем двумя различными способами.

Третий важный момент — *оценка точности*. В большинстве случаев вычисления производятся с приближенными числами и притом приближенно. Поэтому даже для точного метода решения задачи на каждом этапе вычислений возникают *погрешности действий* и

погрешности округлений. Если сам метод — приближенный, то к этим двум погрешностям присоединяется *погрешность метода*. При неблагоприятных обстоятельствах суммарная погрешность может быть столь велика, что полученный результат будет иметь лишь иллюзорное значение. В соответствующих главах книги указаны методы оценки погрешностей для основных вычислений.

В вычислительном бланке полезно предусмотреть столбцы для табличных разностей (см. гл. XIV, § 2), которые можно использовать для контроля вычислений. А именно, если правильность таблицы разностей нарушается на отдельном участке, то следует пересчитать соответствующие элементы таблицы (либо выявить причину нарушения).

Нужно обратить внимание также на *аккуратность* и *четкость* записи в вычислительных бланках. Практика показывает, что нечеткая запись цифр часто приводит к ошибкам и может погубить хорошо организованное вычисление. Особенно опасны ошибки в записи чисел, содержащих большое число нулей. Такие числа следует записывать в нормальной форме, выделяя целую степень десяти, например

$$0{,}00000345 = 3{,}45 \cdot 10^{-6}$$

и т. п.

Дальнейшая часть книги посвящена главным образом методам вычислений. Приводимые числовые примеры во многих случаях упрощены, причем промежуточные выкладки часто опускаются.

ГЛАВА I

ПРИБЛИЖЕННЫЕ ЧИСЛА

§ 1. Абсолютная и относительная погрешности

Приближенным числом a называется число, незначительно отличающееся от точного A и заменяющее последнее в вычислениях. Если известно, что $a < A$, то a называется приближенным значением числа A *по недостатку*; если же $a > A$, то — *по избытку*. Например, для $\sqrt{2}$ число 1,41 будет приближенным значением по недостатку, а 1,42 — по избытку, так как $1{,}41 < \sqrt{2} < 1{,}42$. Если a есть приближенное значение числа A, то пишут $a \approx A$.

Под *ошибкой* или *погрешностью* Δa приближенного числа a обычно понимается разность между соответствующим точным числом A и данным приближенным, т. е.

$$\Delta a = A - a^{*}).$$

Если $A > a$, то ошибка положительна: $\Delta a > 0$; если же $A < a$, то ошибка отрицательна: $\Delta a < 0$. Чтобы получить точное число A, нужно к приближенному числу a прибавить его ошибку Δa, т. е.

$$A = a + \Delta a.$$

Таким образом, точное число можно рассматривать как приближенное с ошибкой, равной нулю.

Во многих случаях знак ошибки неизвестен. Тогда целесообразно пользоваться *абсолютной погрешностью приближенного числа*

$$\Delta = |\Delta a|.$$

Определение 1. *Абсолютной погрешностью* Δ приближенного числа a называется абсолютная величина разности между соответствующим точным числом A и числом a, т. е.

$$\Delta = |A - a|. \tag{1}$$

Здесь следует различать два случая:

1) число A нам известно, тогда абсолютная погрешность Δ легко определяется по формуле (1);

*) Иногда ошибкой называют разность $a - A$.

2) число A нам не известно, что практически бывает чаще всего, и, следовательно, мы не можем определить и абсолютную погрешность Δ по формуле (1).

В этом случае полезно вместо неизвестной теоретической абсолютной погрешности Δ ввести ее оценку сверху, так называемую *предельную абсолютную погрешность.*

Определение 2. Под *предельной абсолютной погрешностью* приближенного числа понимается всякое число, не меньшее абсолютной погрешности этого числа.

Таким образом, если Δ_a — предельная абсолютная погрешность приближенного числа a, заменяющего точное A, то

$$\Delta = |A - a| \leqslant \Delta_a. \tag{2}$$

Отсюда следует, что точное число A заключено в границах

$$a - \Delta_a \leqslant A \leqslant a + \Delta_a. \tag{3}$$

Следовательно, $a - \Delta_a$ есть приближение числа A по недостатку, а $a + \Delta_a$ — приближение числа A по избытку.

В этом случае для краткости пользуются записью

$$A = a \pm \Delta_a.$$

Пример 1. Определить предельную абсолютную погрешность числа $a = 3{,}14$, заменяющего число π.

Решение. Так как имеет место неравенство

$$3{,}14 < \pi < 3{,}15, \quad \text{то} \quad |a - \pi| < 0{,}01$$

и, следовательно, можно принять $\Delta_a = 0{,}01$.

Если учесть, что

$$3{,}14 < \pi < 3{,}142,$$

то будем иметь лучшую оценку: $\Delta_a = 0{,}002$.

Заметим, что сформулированное выше понятие предельной абсолютной погрешности является весьма широким, а именно: *под предельной абсолютной погрешностью приближенного числа a понимается любой представитель бесконечного множества неотрицательных чисел Δ_a, удовлетворяющих неравенству* (2). Отсюда логически вытекает, что всякое число, большее предельной абсолютной погрешности данного приближенного числа, также может быть названо предельной абсолютной погрешностью этого числа. Практически удобно в качестве Δ_a выбирать возможно меньшее при данных обстоятельствах число, удовлетворяющее неравенству (2).

В записи приближенного числа, полученного в результате измерения, обычно отмечают его предельную абсолютную погрешность. Например, если длина отрезка $l = 214$ *см* с точностью до 0,5 *см*, то пишут $l = 214\ см \pm 0{,}5\ см$. Здесь предельная абсолютная погрешность $\Delta_l = 0{,}5\ см$, а точная величина длины l отрезка заключена в границах $213{,}5\ см \leqslant l \leqslant 214{,}5\ см$.

Абсолютная погрешность (или предельная абсолютная погрешность) не достаточна для характеристики точности измерения или вычисления. Так, например, если при измерении длин двух стержней получены результаты $l_1 = 100{,}8\ см \pm 0{,}1\ см$ и $l_2 = 5{,}2\ см \pm 0{,}1\ см$, то, несмотря на совпадение предельных абсолютных погрешностей, качество первого измерения выше, чем второго. Для точности данных измерений существенна абсолютная погрешность, приходящаяся на единицу длины, которая носит название *относительной погрешности.*

Определение 3. *Относительной погрешностью* δ приближенного числа a называется отношение абсолютной погрешности Δ этого числа к модулю соответствующего точного числа $A (A \neq 0)$, т. е.

$$\delta = \frac{\Delta}{|A|}. \tag{4}$$

Отсюда $\Delta = |A| \delta$.

Так же как и для абсолютной погрешности, введем понятие *предельной относительной погрешности.*

Определение 4. *Предельной относительной погрешностью* δ_a данного приближенного числа a называется всякое число, не меньшее относительной погрешности этого числа. По определению имеем:

$$\delta \leqslant \delta_a, \tag{5}$$

т. е. $\frac{\Delta}{|A|} \leqslant \delta_a$, отсюда $\Delta \leqslant |A| \delta_a$.

Таким образом, за предельную абсолютную погрешность числа a можно принять:

$$\Delta_a = |A| \delta_a. \tag{6}$$

Так как на практике $A \approx a$, то вместо формулы (6) часто пользуются формулой

$$\Delta_a = |a| \delta_a. \tag{6'}$$

Отсюда, зная предельную относительную погрешность δ_a, получают границы для точного числа. То обстоятельство, что точное число лежит между $a(1 - \delta_a)$ и $a(1 + \delta_a)$, условно записывают так:

$$A = a(1 \pm \delta_a).$$

Пусть a — приближенное число, заменяющее точное A, и Δ_a — предельная абсолютная погрешность числа a. Положим для определенности, что $A > 0$, $a > 0$ и $\Delta_a < a$. Тогда

$$\delta = \frac{\Delta}{A} \leqslant \frac{\Delta_a}{a - \Delta_a}.$$

Следовательно, в качестве предельной относительной погрешности числа a можно принять число

$$\delta_a = \frac{\Delta_a}{a - \Delta_a}.$$

Аналогично получаем $\Delta = A\delta \leqslant (a + \Delta)\,\delta_a$; отсюда

$$\Delta_a = \frac{a\delta_a}{1 - \delta_a}.$$

Если, как обычно бывает, $\Delta_a \ll a$ и $\delta_a \ll 1$ (знак $\ll$ обозначает «значительно меньше»), то приближенно можно принять:

$$\delta_a \approx \frac{\Delta_a}{a}$$

и

$$\Delta_a \approx a\delta_a.$$

Пример 2. Вес 1 $дм^3$ воды при 0° С $p = 999{,}847\ Г \pm 0{,}001\ Г$. Определить предельную относительную погрешность результата взвешивания.

Решение. Очевидно, что $\Delta_p = 0{,}001\ Г$ и $p \leqslant 999{,}846\ Г$.

Следовательно,

$$\delta_p = \frac{0{,}001}{999{,}846} \approx 10^{-4}\ \%.$$

Пример 3. При определении газовой постоянной для воздуха получили $R = 29{,}25$. Зная, что относительная погрешность этого значения равна $1\ ‰$, найти пределы, в которых заключается R.

Решение. Имеем $\delta_R = 0{,}001$, тогда $\Delta_R = R\delta_R \approx 0{,}03$.

Следовательно, $29{,}22 \leqslant R \leqslant 29{,}28$.

§ 2. Основные источники погрешностей

Погрешности, встречающиеся в математических задачах, могут быть в основном разбиты на пять групп.

1. Погрешности, связанные с самой постановкой математической задачи. Математические формулировки редко точно отображают реальные явления: обычно они дают лишь более или менее идеализированные модели. Как правило, при изучении тех или иных явлений природы мы вынуждены принять некоторые, упрощающие задачу, условия, что вызывает ряд погрешностей (*погрешности задачи*).

Иногда бывает и так, что решить задачу в точной постановке трудно или даже невозможно. Тогда ее заменяют близкой по результатам приближенной задачей. При этом возникает погрешность, которую можно назвать *погрешностью метода*.

2. Погрешности, связанные с наличием бесконечных процессов в математическом анализе. Функции, фигурирующие в математических формулах, часто задаются в виде бесконечных последовательностей или рядов $\left(\text{например, } \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots\right)$. Более того, многие математические уравнения можно решить, лишь описав бесконечные процессы, пределы которых и являются искомыми решениями.

Так как бесконечный процесс, вообще говоря, не может быть завершен в конечное число шагов, то мы вынуждены остановиться на некотором члене последовательности, считая его приближением к искомому решению. Понятно, что такой обрыв процесса вызывает погрешность, называемую обычно *остаточной погрешностью*.

3. Погрешности, связанные с наличием в математических формулах числовых параметров, значения которых могут быть определены лишь приближенно. Таковы, например, все физические константы. Условно назовем эту погрешность *начальной*.

4. Погрешности, связанные с системой счисления. При изображении даже рациональных чисел в десятичной системе или другой позиционной системе справа от запятой может быть бесконечное число цифр (например, может получиться бесконечная десятичная периодическая дробь). При вычислениях, очевидно, можно использовать лишь конечное число этих цифр. Так возникает *погрешность округления*. Например, полагая $\frac{1}{3}=0{,}333$, получаем погрешность $\Delta \approx 3 \cdot 10^{-4}$. Приходится так же округлять и конечные числа, имеющие большое количество знаков.

5. Погрешности, связанные с действиями над приближенными числами (*погрешности действий*). Понятно, что, производя вычисления с приближенными числами, погрешности исходных данных в какой-то мере мы переносим в результат вычислений. В этом отношении погрешности действий являются неустранимыми.

Само собой разумеется, что при решении конкретной задачи те или иные погрешности иногда отсутствуют, или влияние их ничтожно. Но, вообще говоря, для полного анализа погрешностей следует учитывать все их виды. В дальнейшем мы ограничимся в основном исчислением погрешностей действий и погрешностей методов.

§ 3. Десятичная запись приближенных чисел. Значащая цифра. Число верных знаков

Известно, что всякое положительное число a может быть представлено в виде конечной или бесконечной десятичной дроби

$$a=\alpha_m 10^m+\alpha_{m-1}10^{m-1}+\alpha_{m-2}10^{m-2}+\ldots \\ \ldots+\alpha_{m-n+1}10^{m-n+1}+\ldots, \qquad (1)$$

где α_i — цифры числа a ($\alpha_i=0, 1, 2, \ldots, 9$), причем старшая цифра $\alpha_m \neq 0$, а m — некоторое целое число (старший десятичный разряд числа a). Например,

$$3141{,}59\ldots=3\cdot 10^3+1\cdot 10^2+4\cdot 10^1+1\cdot 10^0+ \\ +5\cdot 10^{-1}+9\cdot 10^{-2}+\ldots$$

Каждая единица, стоящая на определенном месте в числе a, написанном в виде десятичной дроби (1), имеет свое значение. Единица, стоящая на первом месте, равна 10^m, на втором — 10^{m-1}, на n-м — 10^{m-n+1} и т. д.

На практике преимущественно приходится иметь дело с приближенными числами, представляющими собой конечные десятичные дроби

$$b = \beta_m 10^m + \beta_{m-1} 10^{m-1} + \ldots + \beta_{m-n+1} 10^{m-n+1} \quad (\beta_m \neq 0). \quad (2)$$

Все сохраняемые десятичные знаки $\beta_i (i = m, m-1, \ldots, m-n+1)$ называются *значащими цифрами* приближенного числа b, причем возможно, что некоторые из них равны нулю (за исключением β_m). При позиционном изображении числа b в десятичной системе счисления иногда приходится вводить лишние нули в начале или в конце числа. Например,

$$b = 7 \cdot 10^{-3} + 0 \cdot 10^{-4} + 1 \cdot 10^{-5} + 0 \cdot 10^{-6} = 0{,}007010,$$

или

$$b = 2 \cdot 10^9 + 0 \cdot 10^8 + 0 \cdot 10^7 + 3 \cdot 10^6 + 0 \cdot 10^5 = 2\,003\,000\,000.$$

Такие нули (в приведенных примерах они подчеркнуты) не считаются значащими цифрами.

Определение 1. *Значащей цифрой* приближенного числа называются всякая цифра в его десятичном изображении, отличная от нуля, и нуль, если он содержится между значащими цифрами или является представителем сохраненного десятичного разряда. Все остальные нули, входящие в состав приближенного числа и служащие лишь для обозначения десятичных разрядов его, не причисляются к значащим цифрам.

Например, в числе 0,002 080 первые три нуля не являются значащими цифрами, так как они служат только для установления десятичных разрядов других цифр. Остальные два нуля являются значащими цифрами, так как первый из них находится между значащими цифрами 2 и 8, а второй, как это отражено в записи, указывает, что в приближенном числе сохранен десятичный разряд 10^{-6}. В случае, если в данном числе 0,002 080 последняя цифра не является значащей, то это число должно быть записано в виде 0,002 08. С этой точки зрения числа 0,002 080 и 0,002 08 не равноценны, так как первое из них содержит четыре значащих цифры, а второе — лишь три значащих цифры.

При написании больших чисел нули справа могут служить как для обозначения значащих цифр, так и для определения разрядов остальных цифр. Поэтому при обычной записи чисел могут возникнуть неясности. Например, рассматривая число 689 000, мы не имеем возможности по его виду судить о том, сколько в нем значащих цифр, хотя можно утверждать, что их не меньше трех. Этой неопределенности можно избежать, выявив десятичный порядок числа

и записав его в виде $6{,}89\cdot 10^5$, если оно имеет три значащих цифры; или $6{,}8900\cdot 10^5$, если число имеет пять значащих цифр, и т. п. Вообще, такого рода запись удобна для чисел, содержащих большое количество незначащих нулей, например $0{,}000\,000\,120 = 1{,}20\cdot 10^{-7}$ и т. п.

Введем понятие о *верных десятичных знаках приближенного числа.*

Определение 2. Говорят, что n первых значащих цифр (десятичных знаков) приближенного числа являются *верными*, если абсолютная погрешность этого числа не превышает половины единицы разряда, выражаемого n-й значащей цифрой, считая слева направо.

Таким образом, если для приближенного числа a (1), заменяющего точное число A, известно, что

$$\Delta = |A - a| \leqslant \frac{1}{2}\cdot 10^{m-n+1},$$

то, по определению, первые n цифр $\alpha_m, \alpha_{m-1}, \ldots, \alpha_{m-n+1}$ этого числа являются верными.

Например, для точного числа $A = 35{,}97$ число $a = 36{,}00$ является приближением с тремя верными знаками, так как $|A - a| = 0{,}03 < \frac{1}{2}\cdot 0{,}1$.

Заметим, что в математических таблицах все помещенные значащие цифры являются верными. Так, например, в пятизначных таблицах логарифмов гарантировано, что абсолютная погрешность мантиссы не превосходит $\frac{1}{2}\cdot 10^{-5}$ и т. п.

Термин «n верных знаков» не следует понимать буквально, т. е. так, что в данном приближенном числе a, имеющем n верных знаков, n первых значащих цифр его совпадают с соответствующими цифрами точного числа A. Например, приближенное число $a = 9{,}995$, заменяющее точное $A = 10$, имеет три верных знака, причем все цифры этих чисел различны. Однако во многих случаях дело обстоит именно так, что верные знаки приближенного числа одинаковы с соответствующими цифрами точного числа.

Замечание. В некоторых случаях удобно говорить, что число a является приближением точного числа A с n *верными знаками* в широком смысле, понимая под этим, что абсолютная погрешность $\Delta = |A - a|$ не превышает единицы десятичного разряда, выражаемого n-й значащей цифрой приближенного числа.

Например, для точного числа $A = 412{,}3567$ число $a = 412{,}356$ является приближением с шестью верными знаками в широком смысле, так как $\Delta = 0{,}0007 < 1\cdot 10^{-3}$.

В дальнейшем верные знаки приближенного числа мы будем понимать в смысле определения 2 (т. е. *в узком смысле*), если явно не оговорено противное.

§ 4. Округление чисел

Рассмотрим некоторое приближенное или точное число a, записанное в десятичной нумерации. Часто бывает надобность в *округлении* этого числа, т. е. в замене его числом a_1 с меньшим количеством значащих цифр. Число a_1 выбирают так, чтобы *погрешность округления* $|a_1 - a|$ была минимальной.

Правило округления (по дополнению). Чтобы округлить число до n значащих цифр, отбрасывают все цифры его, стоящие справа от n-й значащей цифры, или, если это нужно для сохранения разрядов, заменяют их нулями. При этом:

1) если первая из отброшенных цифр меньше 5, то оставшиеся десятичные знаки сохраняются без изменения;

2) если первая из отброшенных цифр больше 5, то к последней оставшейся цифре прибавляется единица;

3) если первая из отброшенных цифр равна 5 и среди остальных отброшенных цифр имеются ненулевые, то последняя оставшаяся цифра увеличивается на единицу;

3а) если же первая из отброшенных цифр равна 5 и все остальные отброшенные цифры являются нулями, то последняя оставшаяся цифра сохраняется неизменной, если она четная, и увеличивается на единицу, если она нечетная (правило четной цифры).

Иными словами, если при округлении числа отбрасывается меньше половины единицы последнего сохраняемого десятичного разряда, то цифры всех сохраненных разрядов остаются неизменными; если же отброшенная часть числа составляет больше половины единицы последнего сохраненного десятичного разряда, то цифра этого разряда увеличивается на единицу. В исключительном случае, когда отброшенная часть в точности равна половине единицы последнего сохраненного десятичного разряда, то для компенсации знаков ошибок округления используется правило четной цифры.

Очевидно, что при применении правила округления погрешность округления не превосходит $\frac{1}{2}$ единицы десятичного разряда, определяемого последней оставленной значащей цифрой.

Пример 1. Округляя число

$$\pi = 3{,}14159\,26535\ldots$$

до пяти, четырех и трех значащих цифр, получим приближенные числа 3,1416; 3,142; 3,14 с абсолютными погрешностями, меньшими $\frac{1}{2}\cdot 10^{-4}$; $\frac{1}{2}\cdot 10^{-3}$ и $\frac{1}{2}\cdot 10^{-2}$.

Пример 2. Округляя число 1,2500 до двух значащих цифр, получим приближенное число 1,2 с абсолютной погрешностью, равной $\frac{1}{2}\cdot 10^{-1} = 0{,}05$.

Точность приближенного числа зависит не от количества значащих цифр, а от количества верных значащих цифр [1], [2]. В тех случаях, когда приближенное число содержит излишнее количество неверных значащих цифр, прибегают к округлению. Обычно руководствуются следующим практическим правилом: *при выполнении приближенных вычислений число значащих цифр промежуточных результатов не должно превышать числа верных цифр более чем на одну или две единицы.* Окончательный результат может содержать не более чем одну излишнюю значащую цифру, по сравнению с верными. Если при этом абсолютная погрешность результата не превышает двух единиц последнего сохраненного десятичного разряда, то излишняя цифра называется *сомнительной.*

Приведенное правило позволяет без ущерба точности вычислений избегать написания лишних цифр и значительно экономит время вычислений. Сохранение запасных знаков имеет тот смысл, что обычно оценка погрешностей результатов производится для наихудших вариантов, и фактическая погрешность может оказаться значительно меньше максимальной теоретической. Таким образом, во многих случаях те значащие цифры, которые считаются неверными, на самом деле являются верными.

Приходится также округлять точные числа, содержащие слишком много или бесконечное количество значащих цифр, сообразуясь с общей точностью вычислений.

Заметим, что если точное число A округлить по правилу дополнения до n значащих цифр, то полученное таким образом приближенное число a будет иметь n верных цифр (в узком смысле).

Если же *приближенное число* a, имеющее n верных цифр, округлить до n значащих цифр, то полученное новое приближенное число a_1, вообще говоря, будет иметь n верных цифр в широком смысле. Действительно, в силу неравенства

$$|A-a_1| \leqslant |A-a|+|a-a_1|$$

предельная абсолютная погрешность числа a_1 складывается из абсолютной погрешности числа a и погрешности округления.

§ 5. Связь относительной погрешности приближенного числа с количеством верных знаков этого числа

Докажем теорему, которая связывает величину относительной погрешности приближенного числа с количеством верных знаков этого числа [3], [4].

Теорема. *Если положительное приближенное число a имеет n верных десятичных знаков в узком смысле, то относительная погрешность δ этого числа не превосходит* $\left(\frac{1}{10}\right)^{n-1}$, *деленную на первую*

значащую цифру данного числа, т. е.

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1},$$

где α_m — первая значащая цифра числа a.

Доказательство. Пусть число

$$a = \alpha_m 10^m + \alpha_{m-1}10^{m-1} + \ldots + \alpha_{m-n+1}10^{m-n+1} + \ldots \quad (\alpha_m \geqslant 1)$$

является приближенным значением точного числа A и имеет n верных знаков. Тогда по определению имеем:

$$\Delta = |A - a| \leqslant \frac{1}{2} \cdot 10^{m-n+1};$$

отсюда

$$A \geqslant a - \frac{1}{2} \cdot 10^{m-n+1}.$$

Последнее неравенство еще более усилится, если число a заменим заведомо меньшим числом $\alpha_m 10^m$,

$$A \geqslant \alpha_m 10^m - \frac{1}{2} \cdot 10^{m-n+1} = \frac{1}{2} \cdot 10^m \left(2\alpha_m - \frac{1}{10^{n-1}}\right). \qquad (1)$$

Правая часть неравенства (1) достигает наименьшего значения при $n = 1$. Поэтому

$$A \geqslant \frac{1}{2} \cdot 10^m (2\alpha_m - 1), \qquad (2)$$

или, так как

$$2\alpha_m - 1 = \alpha_m + (\alpha_m - 1) \geqslant \alpha_m,$$

то

$$A \geqslant \frac{1}{2}\alpha_m 10^m.$$

Следовательно,

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2}10^{m-n+1}}{\frac{1}{2}\alpha_m 10^m} = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Итак,

$$\delta \leqslant \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \qquad (3)$$

Теорема доказана.

Замечание 1. Пользуясь неравенством (2), можно получить более точную оценку относительной погрешности δ.

Следствие 1. За предельную относительную погрешность числа a можно принять:

$$\delta_a = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}, \tag{4}$$

где α_m — первая значащая цифра числа a.

Следствие 2. Если число a имеет больше двух верных знаков, т. е. $n \geqslant 2$, то практически справедлива формула

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \tag{5}$$

Действительно, при $n \geqslant 2$ числом $\frac{1}{10^{n-1}}$ в неравенстве (1) можно пренебречь. Тогда

$$A \geqslant \frac{1}{2} \cdot 10^m \cdot 2\alpha_m = \alpha_m 10^m;$$

отсюда

$$\delta = \frac{\Delta}{A} \leqslant \frac{\frac{1}{2} \cdot 10^{m-n+1}}{\alpha_m 10^m} = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Следовательно,

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Замечание 2. Если приближенное число a имеет n верных десятичных знаков в широком смысле, то оценки (4) и (5) следует увеличить в два раза.

Пример 1. Какова предельная относительная погрешность, если вместо числа π взять число $a = 3{,}14$?

Решение. В нашем случае $\alpha_m = 3$ и $n = 3$. Следовательно,

$$\delta_a = \frac{1}{2 \cdot 3}\left(\frac{1}{10}\right)^{3-1} = \frac{1}{600} = \frac{1}{6}\%.$$

Пример 2. Со сколькими десятичными знаками надо взять $\sqrt{20}$, чтобы погрешность не превышала 0,1%?

Решение. Так как первая цифра 4, то $\alpha_m = 4$, причем $\delta = 0{,}001$. Имеем $\frac{1}{4 \cdot 10^{n-1}} \leqslant 0{,}001$, отсюда $10^{n-1} \geqslant 250$ и $n \geqslant 4$.

Приведенная теорема дает возможность по числу верных знаков приближенного числа

$$a = \alpha_m \cdot 10^m + \alpha_{m-1} 10^{m-1} + \ldots \tag{6}$$

определить его относительную погрешность δ.

Для решения обратной задачи — определения количества n верных знаков числа (6), если известна его относительная погрешность δ, обычно пользуются приближенной формулой

$$\delta = \frac{\Delta}{a} \quad (a > 0),$$

где Δ — абсолютная погрешность числа a. Отсюда

$$\Delta = a\delta. \tag{7}$$

Учитывая старший десятичный разряд числа Δ, легко установить количество верных знаков данного приближенного числа a. В частности, если

$$\delta \leqslant \frac{1}{10^n},$$

то из формул (6) и (7) имеем:

$$\Delta \leqslant (\alpha_m + 1)\, 10^m \cdot 10^{-n} \leqslant 10^{m-n+1},$$

т. е. число a заведомо имеет n верных десятичных знаков в широком смысле. Аналогично, если

$$\delta \leqslant \frac{1}{2 \cdot 10^n},$$

то число a имеет n верных знаков в узком смысле.

Пример 3. Приближенное число $a = 24\,253$ имеет относительную точность $1\,\%$. Сколько в нем верных знаков?

Решение. Имеем:

$$\Delta = 24\,253 \cdot 0{,}01 \approx 243 = 2{,}43 \cdot 10^2.$$

Следовательно, число a имеет *верными лишь* первые две цифры ($n = 2$); цифра сотен является сомнительной. Согласно приведенному выше правилу число a предпочтительнее записать в виде $a = 2{,}43 \cdot 10^4$.

Замечание. Указанный способ определения числа верных знаков является приближенным. При точном подсчете верных цифр числа a следует исходить из неравенств

$$\delta \geqslant \frac{\Delta}{a + \Delta}$$

и

$$\Delta \leqslant \frac{a\delta}{1 - \delta} \quad (0 \leqslant \delta < 1).$$

§ 6. Таблицы для определения предельной относительной погрешности по числу верных знаков и наоборот

Если приближенное число написано с указанными верными десятичными знаками, то можно легко подсчитать его предельную относительную погрешность. Практически с таким подсчетом приходится сталкиваться часто и поэтому желательно рационализиро-

вать эту операцию. Таблица 2 [5] указывает относительную погрешность в процентах приближенного числа в зависимости от количества верных в широком смысле десятичных знаков его и от первых двух значащих цифр числа, считая слева направо.

Таблица 2

Относительная погрешность (в %) чисел с n верными знаками

Первые две значащие цифры	n		
	2	3	4
10—11	10	1	0,1
12—13	8,3	0,83	0,083
14, ..., 16	7,1	0,71	0,071
17, ..., 19	5,9	0,59	0,059
20, ..., 22	5	0,5	0,05
23, ..., 25	4,3	0,43	0,043
26, ..., 29	3,8	0,38	0,038
30, ..., 34	3,3	0,33	0,033
35, ..., 39	2,9	0,29	0,029
40, ..., 44	2,5	0,25	0,025
45, ..., 49	2,2	0,22	0,022
50, ..., 59	2	0,2	0,02
60, ..., 69	1,7	0,17	0,017
70, ..., 79	1,4	0,14	0,014
80, ..., 89	1,2	0,12	0,012
90, ..., 99	1,1	0,11	0,011

Пусть, например, имеем приближенное число 0,00354 с тремя верными десятичными знаками. Так как здесь $n=3$ и число 35 содержится в промежутке 35, ..., 39, то по таблице 2 находим $\delta=0{,}29\ \%$.

Если известна только первая цифра числа, например 4, то берем, конечно, большее из чисел 2,5 и 2,2, соответствующих возможным вариантам 40, ..., 44 и 45, ..., 49 (при $n=2$). Если и первая цифра неизвестна, то берем числа из первой строки (10 %; 1 %; 0,1 %), как наибольшие. Из этой таблицы мы видим, что три верных знака обеспечивают относительную точность (не менее 1 %), достаточную для большинства технических расчетов. Заметим, что если приближенное число имеет два, три или четыре верных знака в узком смысле, то все числа таблицы нужно уменьшить вдвое.

В таблице 3 [5] приведены верхние границы для относительных погрешностей (в процентах), обеспечивающих данному приближенному значению то или другое число верных знаков в широком смысле в зависимости от его первых двух цифр.

Таблица 3

Число верных знаков приближенного числа в зависимости от предельной относительной погрешности (в %)

Первые две значащие цифры	n		
	2	3	4
10—11	4,2	0,42	0,042
12—13	3,6	0,36	0,036
14, ..., 16	2,9	0,29	0,029
17, ..., 19	2,5	0,25	0,025
20, ..., 22	2,2	0,22	0,022
23, ..., 25	1,9	0,19	0,019
26, ..., 29	1,7	0,17	0,017
30, ..., 34	1,4	0,14	0,014
35, ..., 39	1,2	0,12	0,012
40, ..., 44	1,1	0,11	0,011
45, ..., 49	1	0,1	0,01
50, ..., 54	0,9	0,09	0,009
55, ..., 59	0,8	0,08	0,008
60, ..., 69	0,7	0,07	0,007
70, ..., 79	0,6	0,06	0,006
80, ..., 99	0,5	0,05	0,005

Покажем на примере, как надо пользоваться таблицей 3. Пусть, например, дано приближенное число $a = 5{,}297$ с относительной погрешностью $\delta = 0{,}5\%$. Здесь первые две значащие цифры 5 и 2; число, образованное этими цифрами, содержится между 50 и 54, причем последним, в зависимости от числа верных знаков, соответствуют относительные погрешности $0{,}9\%$; $0{,}09\%$; $0{,}009\%$ и т. д. Так как $\delta = 0{,}5\% < 0{,}9\%$ и относительная погрешность числа не зависит от того, какие десятичные разряды выражают цифры этого числа, то число $a = 5{,}297$ имеет два верных десятичных знака в широком смысле.

Примеры. 1. Полагая $\pi = 3{,}142$; $\sqrt{7} = 2{,}65$; $e = 2{,}718$; $\lg 5 = 0{,}699$; $\sin 1^\circ = 0{,}0174$, по таблице 2 находим, что соответствующие относительные погрешности следующие: $\delta = 0{,}033\%$; $\delta = 0{,}19\%$; $\delta = 0{,}019\%$; $\delta = 0{,}17\%$; $\delta = 0{,}59\%$.

2. По прогибу стального стержня вычислен модуль Юнга $E = 2212...$ $т/см^2$ с точностью до 2%. Сколько верных знаков в найденном значении? По таблице 3 находим $n = 2$. Следовательно, $E = 22 \cdot 10^2$ $т/см^2$.

3. Для взрывчатой смеси в газомоторе вычислена газовая постоянная $R = 31{,}5...$ с относительной погрешностью $\delta = 1\%$. Определить число верных знаков. По таблице 3 находим $n = 2$. Значит, $R = 32$.

§ 7. Погрешность суммы

Теорема 1. *Абсолютная погрешность алгебраической суммы нескольких приближенных чисел не превышает суммы абсолютных погрешностей этих чисел.*

Доказательство. Пусть x_1, x_2, ..., x_n — данные приближенные числа. Рассмотрим их алгебраическую сумму

$$u = \pm x_1 \pm x_2 \pm \ldots \pm x_n.$$

Очевидно, что

$$\Delta u = \pm \Delta x_1 \pm \Delta x_2 \pm \ldots \pm \Delta x_n$$

и, следовательно,

$$|\Delta u| \leqslant |\Delta x_1| + |\Delta x_2| + \ldots + |\Delta x_n|. \tag{1}$$

Следствие. За предельную абсолютную погрешность алгебраической суммы можно принять сумму предельных абсолютных погрешностей слагаемых

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}. \tag{2}$$

Из формулы (2) следует, что предельная абсолютная погрешность суммы не может быть меньше предельной абсолютной погрешности наименее точного (в смысле абсолютной погрешности) из слагаемых, т. е. слагаемого, имеющего максимальную абсолютную погрешность. Следовательно, с какой бы степенью точности ни были определены остальные слагаемые, мы не можем за их счет увеличить точность суммы. Поэтому не имеет смысла сохранять излишние знаки и в более точных слагаемых. Отсюда вытекает следующее, обычно применяемое, практическое правило для сложения приближенных чисел.

Правило. Чтобы сложить числа различной абсолютной точности, следует:

1) выделить числа, десятичная запись которых обрывается ранее других, и оставить их без изменения;

2) остальные числа округлить по образцу выделенных, сохраняя один или два запасных десятичных знака;

3) произвести сложение данных чисел, учитывая все сохраненные знаки;

4) полученный результат округлить на один знак.

При округлении по правилу дополнения слагаемых суммы

$$u = x_1 + x_2 + \ldots + x_n$$

до m-го десятичного разряда погрешность округления суммы в самом неблагоприятном случае не превышает величины

$$\Delta_{\text{окр}} \leqslant n \cdot \frac{1}{2} \cdot 10^m. \tag{3}$$

Можно получить более точный расчет погрешности округления суммы, если учесть знаки ошибок округления слагаемых.

Пример. Найти сумму приближенных чисел: 0,348; 0,1834; 345,4; 235,2; 11,75; 9,27; 0,0849; 0,0214; 0,000354, каждое из которых имеет все верные значащие цифры (в широком смысле).

Решение. Выделяем числа наименьшей точности 345,4 и 235,2, абсолютная погрешность которых может достигать 0,1. Округляя остальные числа с точностью до 0,01, получим:

$$\begin{array}{r} 345{,}4 \\ 235{,}2 \\ 11{,}75 \\ 9{,}27 \\ 0{,}35 \\ 0{,}18 \\ 0{,}08 \\ 0{,}02 \\ 0{,}00 \\ \hline 602{,}25 \end{array}$$

Округляя результат до 0,1 по правилу четной цифры, получим приближенное значение суммы 602,2.

Полная погрешность Δ результата складывается из трех слагаемых:

1) суммы предельных погрешностей исходных данных

$$\Delta_1 = 10^{-3} + 10^{-4} + 10^{-1} + 10^{-1} + 10^{-2} + 10^{-2} + 10^{-4} + 10^{-4} + 10^{-6} = \\ = 0{,}221\,301 < 0{,}222;$$

2) абсолютной величины суммы ошибок (с учетом их знаков) округления слагаемых

$$\Delta_2 = |-0{,}002 + 0{,}0034 + 0{,}0049 + 0{,}0014 + 0{,}000\,354| = \\ = 0{,}008054 < 0{,}009;$$

3) заключительной погрешности округления результата

$$\Delta_3 = 0{,}050.$$

Следовательно,

$$\Delta = \Delta_1 + \Delta_2 + \Delta_3 \leqslant 0{,}222 + 0{,}009 + 0{,}050 = 0{,}281 < 0{,}3;$$

и, таким образом, искомая сумма есть $602{,}2 \pm 0{,}3$.

Теорема 2. *Если слагаемые — одного и того же знака, то предельная относительная погрешность их суммы не превышает наибольшей из предельных относительных погрешностей слагаемых.*

Доказательство. Пусть $u = x_1 + x_2 + \ldots + x_n$, где, для определенности, $x_i > 0$ $(i = 1, 2, \ldots, n)$.

Обозначим через A_i ($A_i > 0$; $i = 1, 2, \ldots, n$) точные величины слагаемых x_i, а через $A = A_1 + A_2 + \ldots + A_n$ — точное значение суммы u. Тогда за предельную относительную погрешность суммы можно принять:

$$\delta_u = \frac{\Delta_u}{A} = \frac{\Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}}{A_1 + A_2 + \ldots + A_n}. \tag{4}$$

Так как

$$\delta_{x_i} = \frac{\Delta_{x_i}}{A_i} \quad (i = 1, 2, \ldots, n),$$

то

$$\Delta_{x_i} = A_i \delta_{x_i}. \tag{4'}$$

Подставляя это выражение в формулу (4), получим:

$$\delta_u = \frac{A_1\delta_{x_1} + A_2\delta_{x_2} + \ldots + A_n\delta_{x_n}}{A_1 + A_2 + \ldots + A_n}.$$

Пусть $\bar{\delta}$ является наибольшей из относительных погрешностей δ_{x_i}, т. е. $\delta_{x_i} \leqslant \bar{\delta}$. Тогда

$$\delta_u \leqslant \frac{\bar{\delta}(A_1 + A_2 + \ldots + A_n)}{A_1 + A_2 + \ldots + A_n} = \bar{\delta}.$$

Следовательно, $\delta_u \leqslant \bar{\delta}$, т. е.

$$\delta_u \leqslant \max(\delta_{x_1}, \delta_{x_2}, \ldots, \delta_{x_n}).$$

§ 8. Погрешность разности

Рассмотрим разность двух приближенных чисел $u = x_1 - x_2$.

По формуле (2) § 7 предельная абсолютная погрешность Δ_u разности

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2},$$

т. е. *предельная абсолютная погрешность разности равна сумме предельных абсолютных погрешностей уменьшаемого и вычитаемого.*

Отсюда предельная относительная погрешность разности

$$\delta_u = \frac{\Delta_{x_1} + \Delta_{x_2}}{A}, \tag{1}$$

где A — точное значение абсолютной величины разности чисел x_1 и x_2.

Замечание о потере точности при вычитании близких чисел. Если приближенные числа x_1 и x_2 достаточно близки друг к другу и имеют малые абсолютные погрешности, то число A мало. Из формулы (1) вытекает, что предельная относительная погрешность в этом случае может быть весьма большой, в то

время как относительные погрешности уменьшаемого и вычитаемого остаются малыми, т. е. здесь происходит *потеря точности.*

Вычислим, например, разность двух чисел: $x_1 = 47{,}132$ и $x_2 = 47{,}111$, каждое из которых имеет пять верных значащих цифр. Вычитая, получим $u = 47{,}132 - 47{,}111 = 0{,}021$.

Таким образом, разность u имеет лишь две значащие цифры, из которых последняя сомнительна, так как предельная абсолютная погрешность разности

$$\Delta_u = 0{,}0005 + 0{,}0005 = 0{,}001.$$

Предельные относительные погрешности вычитаемого, уменьшаемого и разности соответственно

$$\delta_{x_1} = \frac{0{,}0005}{47{,}132} \approx 0{,}00001;$$

$$\delta_{x_2} = \frac{0{,}0005}{47{,}111} \approx 0{,}00001;$$

$$\delta_u = \frac{0{,}001}{0{,}021} \approx 0{,}05.$$

Предельная относительная погрешность разности здесь примерно в 5000 раз больше предельных относительных погрешностей исходных данных.

Поэтому при приближенных вычислениях полезно преобразовывать выражения, вычисление числовых значений которых приводит к вычитанию близких чисел.

Пример. Найти разность

$$u = \sqrt{2{,}01} - \sqrt{2} \tag{2}$$

с тремя верными знаками.

Решение. Так как

$$\sqrt{2{,}01} = 1{,}4177\,4469\ldots$$

и

$$\sqrt{2} = 1{,}4142\,1356\ldots,$$

то искомый результат есть

$$u = 0{,}00353 = 3{,}53 \cdot 10^{-3}.$$

Этот результат можно получить, если записать выражение (2) в виде

$$u = \frac{0{,}01}{\sqrt{2{,}01} + \sqrt{2}}$$

и взять корни $\sqrt{2{,}01}$ и $\sqrt{2}$ лишь с тремя верными знаками. Действительно,

$$u = \frac{0{,}01}{1{,}42 + 1{,}41} = \frac{0{,}01}{2{,}83} = 10^{-2} \cdot 3{,}53 \cdot 10^{-1} = 3{,}53 \cdot 10^{-3}.$$

Исходя из вышесказанного, получаем следующее практическое правило: при приближенных вычислениях следует по возможности избегать вычитания двух почти равных приближенных чисел; если же в силу необходимости приходится вычитать такие числа, то следует уменьшаемое и вычитаемое брать с достаточным числом запасных верных знаков (если такая возможность имеется). Например, если известно, что при вычитании чисел x_1 и x_2 первые m значащих цифр их пропадут, а результат необходимо иметь с n верными значащими цифрами, то следует взять x_1 и x_2 с $m+n$ верными значащими цифрами.

§ 9. Погрешность произведения

Теорема. *Относительная погрешность произведения нескольких приближенных чисел, отличных от нуля, не превышает суммы относительных погрешностей этих чисел.*

Доказательство. Пусть $u = x_1 x_2 \dots x_n$.

Предполагая для простоты, что приближенные числа $x_1, x_2, \dots, x_n$ положительны, будем иметь:

$$\ln u = \ln x_1 + \ln x_2 + \dots + \ln x_n.$$

Отсюда, используя приближенную формулу $\Delta \ln x \approx d \ln x = \frac{\Delta x}{x}$, находим:

$$\frac{\Delta u}{u} = \frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2} + \dots + \frac{\Delta x_n}{x_n}.$$

Оценивая последнее выражение по абсолютной величине, получим:

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x_1}{x_1}\right| + \left|\frac{\Delta x_2}{x_2}\right| + \dots + \left|\frac{\Delta x_n}{x_n}\right|.$$

Если A_i $(i = 1, 2, \dots, n)$ — точные значения сомножителей x_i и $|\Delta x_i|$, как это бывает обычно, малы по сравнению с x_i, то приближенно можно положить:

$$\left|\frac{\Delta x_i}{x_i}\right| \approx \left|\frac{\Delta x_i}{A_i}\right| = \delta_i$$

и

$$\left|\frac{\Delta u}{u}\right| = \delta,$$

где δ_i — относительные погрешности сомножителей x_i $(i = 1, 2, \dots, n)$ и δ — относительная погрешность произведения.

Следовательно,

$$\delta \leqslant \delta_1 + \delta_2 + \dots + \delta_n. \qquad (1)$$

Формула (1), очевидно, остается верной также, если сомножители x_i $(i = 1, 2, \dots, n)$ имеют различные знаки.

Следствие. Предельная относительная погрешность произведения равна сумме предельных относительных погрешностей сомножителей, т. е.

$$\delta_u = \delta_{x_1} + \delta_{x_2} + \ldots + \delta_{x_n}. \qquad (2)$$

Если все множители произведения u весьма точны, за исключением одного, то из формулы (2) следует, что предельная относительная погрешность произведения в этом случае будет практически совпадать с предельной относительной погрешностью множителя, обладающего наименьшей точностью. В частном случае, если приближенным является лишь множитель x_1, то имеем просто

$$\delta_u = \delta_{x_1}.$$

Зная предельную относительную погрешность δ_u произведения u, можно определить его предельную абсолютную погрешность Δ_u по формуле

$$\Delta_u = |u|\,\delta_u.$$

Пример 1. Определить произведение u приближенных чисел $x_1 = 12{,}2$ и $x_2 = 73{,}56$ и число верных знаков в нем, если все написанные цифры сомножителей верные.

Решение. Имеем $\Delta_{x_1} = 0{,}05$ и $\Delta_{x_2} = 0{,}005$. Отсюда

$$\delta_u = \frac{0{,}05}{12{,}2} + \frac{0{,}005}{73{,}56} = 0{,}0042.$$

Так как произведение $u = 897{,}432$, то $\Delta_u = u\delta_u = 897 \cdot 0{,}004 = 3{,}6$ (приблизительно).

Отсюда u имеет лишь два верных знака и результат следует записать так:

$$u = 897 \pm 4.$$

Отметим частный случай

$$u = kx,$$

где k — точный множитель, отличный от нуля. Имеем:

$$\delta_u = \delta_x$$

и

$$\Delta_u = |k|\,\Delta_x,$$

т. е. *при умножении приближенного числа на точный множитель k относительная предельная погрешность не изменяется, а абсолютная предельная погрешность увеличивается в $|k|$ раз.*

Пример 2. При наведении ракеты на цель предельная угловая ошибка $\varepsilon = 1'$. Каково возможное отклонение Δ_u ракеты от цели на дальности $x = 2000$ *км* при отсутствии корректирования ошибки?

Решение. Здесь

$$\Delta_u = \frac{\pi}{180 \cdot 60} \cdot 2000 \text{ км} \approx 580 \text{ м}.$$

Очевидно, что относительная погрешность произведения не может быть меньше, чем относительная погрешность наименее точного из сомножителей. Поэтому здесь, как и в случае сложения, не имеет смысла сохранять в более точных сомножителях излишнее количество значащих цифр.

Полезно руководствоваться следующим правилом: чтобы найти произведение нескольких приближенных чисел с различным числом верных значащих цифр, достаточно:

1) округлить их так, чтобы каждое из них содержало на одну (или две) значащую цифру больше, чем число верных цифр в наименее точном из сомножителей;

2) в результате умножения сохранить столько значащих цифр, сколько верных цифр имеется в наименее точном из сомножителей (или удержать еще одну запасную цифру).

Пример 3. Найти произведение приближенных чисел $x_1 = 2{,}5$ и $x_2 = 72{,}397$, верных в написанных знаках.

Решение. Применяя правило, после округления имеем $x_1 = 2{,}5$ и $x_2 = 72{,}4$. Отсюда $x_1 x_2 = 2{,}5 \cdot 72{,}4 = 181 \approx 1{,}8 \cdot 10^2$.

§ 10. Число верных знаков произведения

Пусть имеем произведение n сомножителей ($n \leqslant 10$) $u = x_1 x_2 \ldots x_n$ каждый из которых имеет по крайней мере m ($m > 1$) верных цифр. Пусть, далее, $\alpha_1, \alpha_2, \ldots, \alpha_n$ — первые значащие цифры в десятичной записи множителей:

$$x_i = \alpha_i 10^{p_i} + \beta_i 10^{p_i - 1} + \ldots \quad (i = 1, 2, 3, \ldots, n).$$

Тогда по формуле (5) § 5 будем иметь:

$$\delta_{x_i} = \frac{1}{2\alpha_i} \left(\frac{1}{10}\right)^{m-1} \quad (i = 1, 2, \ldots, n)$$

и, следовательно,

$$\delta_u = \frac{1}{2} \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n}\right) \left(\frac{1}{10}\right)^{m-1}. \tag{1}$$

Так как $\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_n} \leqslant 10$, то $\delta_u \leqslant \frac{1}{2} \left(\frac{1}{10}\right)^{m-2}$.

Следовательно, в самом неблагоприятном случае произведение u имеет $m - 2$ верных знака.

Правило. Если все сомножители имеют m верных десятичных знаков и число их не больше 10, то число верных (в широком смысле) знаков произведения на одну или на две единицы меньше m.

Следовательно, если нужно обеспечить в произведении m верных десятичных знаков, то сомножители следует брать с одним или двумя запасными знаками.

Если сомножители обладают различной точностью, то под m следует понимать число верных знаков в наименее точном из сомножителей. Таким образом, *число верных знаков произведения небольшого числа сомножителей (порядка десяти) может быть на одну или две единицы меньше числа верных знаков в наименее точном из этих сомножителей.*

Пример 1. Определить относительную погрешность и количество верных цифр произведения $u = 93{,}87 \cdot 9{,}236$.

Решение. По формуле (1) имеем:

$$\delta_u = \frac{1}{2}\left(\frac{1}{9} + \frac{1}{9}\right)\frac{1}{10^3} = \frac{1}{9}\cdot 10^{-3} < \frac{1}{2}\cdot 10^{-3}.$$

Следовательно, произведение u имеет по меньшей мере три верные цифры (см. § 5).

Пример 2. Определить относительную погрешность и число верных цифр произведения $u = 17{,}63 \cdot 14{,}285$.

Решение.

$$\delta_u = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{1}\right)\frac{1}{10^3} = 1\cdot 10^{-3}.$$

Следовательно, в произведении будут по крайней мере три верные цифры (в широком смысле).

§ 11. Погрешность частного

Если $u = \frac{x}{y}$, то $\ln u = \ln x - \ln y$

и

$$\frac{\Delta u}{u} = \frac{\Delta x}{x} - \frac{\Delta y}{y}.$$

Отсюда

$$\left|\frac{\Delta u}{u}\right| \leqslant \left|\frac{\Delta x}{x}\right| + \left|\frac{\Delta y}{y}\right|.$$

Из последней формулы вытекает, что теорема § 9 верна и для частного.

Теорема. *Относительная погрешность частного не превышает суммы относительных погрешностей делимого и делителя.*

Следствие. Если $u = \frac{x}{y}$, то $\delta_u = \delta_x + \delta_y$.

Пример. Найти число верных знаков частного $u = 25{,}7 : 3{,}6$, если все написанные знаки делимого и делителя верны.

Решение. Имеем:

$$\delta_u = \frac{0{,}05}{25{,}7} + \frac{0{,}05}{3{,}6} = 0{,}002 + 0{,}014 = 0{,}016.$$

Так как $u=7{,}14$, то $\Delta_u=0{,}016\cdot 7{,}14=0{,}11$. Поэтому частное u имеет два верных знака в широком смысле, т. е. $u=7{,}1$ или, более точно,

$$u=7{,}14\pm 0{,}11.$$

§ 12. Число верных знаков частного

Пусть делимое x и делитель y имеют по меньшей мере m верных цифр. Если α и β — их первые значащие цифры, то за предельную относительную погрешность частного u может быть принята величина

$$\delta_u=\frac{1}{2}\left(\frac{1}{\alpha}+\frac{1}{\beta}\right)\left(\frac{1}{10}\right)^{m-1}.$$

Отсюда получаем правило: 1) если $\alpha\geqslant 2$ и $\beta\geqslant 2$, то частное u имеет по меньшей мере $m-1$ верных знаков; 2) если $\alpha=1$ или $\beta=1$, то частное u заведомо имеет $m-2$ верных знака.

§ 13. Относительная погрешность степени

Пусть $u=x^m$ (m — натуральное число), тогда $\ln u=m\ln x$ и, следовательно,

$$\left|\frac{\Delta u}{u}\right|=m\left|\frac{\Delta x}{x}\right|.$$

Отсюда

$$\delta_u=m\delta_x, \tag{1}$$

т. е. предельная относительная погрешность m-й степени числа в m раз больше предельной относительной погрешности самого числа.

§ 14. Относительная погрешность корня

Пусть теперь $u=\sqrt[m]{x}$, тогда $u^m=x$. Отсюда

$$\delta_u=\frac{1}{m}\,\delta_x, \tag{1}$$

т. е. предельная относительная погрешность корня m-й степени в m раз меньше предельной относительной погрешности подкоренного числа.

Пример. Определить, с какой относительной погрешностью и со сколькими верными цифрами можно найти сторону a квадрата, если его площадь $s=12{,}34$ (с точностью до 0,01).

Решение. Имеем $a=\sqrt{s}=3{,}5128\ldots$ Так как

$$\delta_s=\frac{0{,}01}{12{,}33}\approx 0{,}0008,$$

то $\delta_a = \frac{1}{2}\delta_s = 0{,}0004$. Поэтому

$$\Delta_a = 3{,}5128 \cdot 0{,}0004 = 1{,}4 \cdot 10^{-3}.$$

Отсюда число a будет иметь примерно четыре верных знака (в широком смысле) и, следовательно, $a = 3{,}513$.

§ 15. Вычисления без точного учета погрешностей

В предыдущих параграфах мы указали способы оценки предельной абсолютной погрешности действий. При этом предполагалось, что абсолютные погрешности компонент усиливают друг друга, что практически бывает сравнительно редко.

При массовых вычислениях, когда не учитывают погрешность каждого отдельного результата, рекомендуется пользоваться следующими правилами подсчета цифр [6].

1. При сложении и вычитании приближенных чисел младший сохраненный десятичный разряд результата должен являться наибольшим среди десятичных разрядов, выражаемых последними верными значащими цифрами исходных данных.

2. При умножении и делении приближенных чисел в результате следует сохранять столько значащих цифр, сколько их имеет приближенное данное с наименьшим числом верных значащих цифр.

3. При возведении в квадрат или куб приближенного числа в результате нужно сохранять столько значащих цифр, сколько верных значащих цифр имеет основание степени.

4. При извлечении квадратного и кубического корней из приближенного числа в результате следует брать столько значащих цифр, сколько верных цифр имеет подкоренное число.

5. Во всех промежуточных результатах следует сохранять на одну цифру больше, чем рекомендуют предыдущие правила. В окончательном результате эта «запасная цифра» отбрасывается.

6. При вычислениях с помощью логарифмов рекомендуется подсчитать число верных значащих цифр в приближенном числе, имеющем наименьшее число верных значащих цифр, и воспользоваться таблицей логарифмов с числом десятичных знаков, на единицу бо́льшим. В окончательном результате последняя значащая цифра отбрасывается.

7. Если данные можно брать с произвольной точностью, то для получения результата с k верными цифрами исходные данные следует брать с таким числом цифр, которые согласно предыдущим правилам обеспечивают $k+1$ верную цифру в результате.

Если некоторые данные имеют излишние младшие десятичные разряды (при сложении и вычитании) или больше значащих цифр, чем другие (при умножении, делении, возведении в степень, извлечении корня), то их предварительно нужно округлить, сохраняя одну запасную цифру.

§ 16. Общая формула для погрешности

Основная задача теории погрешности заключается в следующем: известны погрешности некоторой системы величин, требуется определить погрешность данной функции от этих величин.

Пусть задана дифференцируемая функция

$$u = f(x_1, x_2, \ldots, x_n)$$

и пусть $|\Delta x_i|\ (i = 1, 2, \ldots, n)$ — абсолютные погрешности аргументов функции. Тогда абсолютная погрешность функции

$$|\Delta u| = |f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) - f(x_1, x_2, \ldots, x_n)|.$$

Обычно на практике $|\Delta x_i|$ — малые величины, произведениями, квадратами и высшими степенями которых можно пренебречь. Поэтому можно положить:

$$|\Delta u| \approx |df(x_1, x_2, \ldots, x_n)| = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\Delta x_i\right| \leqslant \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right||\Delta x_i|.$$

Итак,

$$|\Delta u| \leqslant \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right||\Delta x_i|. \tag{1}$$

Отсюда, обозначая через $\Delta_{x_i}\ (i = 1, 2, \ldots, n)$ предельные абсолютные погрешности аргументов x_i и через Δ_u — предельную погрешность функции u, для малых Δx_i получим:

$$\Delta_u = \sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}. \tag{2}$$

Разделив обе части неравенства (1) на u, будем иметь оценку для относительной погрешности функции u

$$\delta \leqslant \sum_{i=1}^{n}\left|\frac{\frac{\partial f}{\partial x_i}}{u}\right||\Delta x_i| = \sum_{i=1}^{n}\left|\frac{\partial}{\partial x_i}\ln f(x_1, \ldots, x_n)\right||\Delta x_i|. \tag{3}$$

Следовательно, за предельную относительную погрешность функции u можно принять:

$$\delta_u = \sum_{i=1}^{n}\left|\frac{\partial}{\partial x_i}\ln u\right|\Delta_{x_i}. \tag{4}$$

Пример 1. Найти предельные абсолютную и относительную погрешности объема шара $V = \frac{1}{6}\pi d^3$, если диаметр $d = 3{,}7\ см \pm 0{,}05\ см$, а $\pi \approx 3{,}14$.

Решение. Рассматривая π и d как переменные величины, вычисляем частные производные

$$\frac{\partial V}{\partial \pi} = \frac{1}{6} d^3 = 8{,}44;$$

$$\frac{\partial V}{\partial d} = \frac{1}{2} \pi d^2 = 21{,}5.$$

В силу формулы (2) предельная абсолютная погрешность объема

$$\Delta V = \left|\frac{\partial V}{\partial \pi}\right| |\Delta \pi| + \left|\frac{\partial V}{\partial d}\right| |\Delta d| = 8{,}44 \cdot 0{,}0016 + 21{,}5 \cdot 0{,}05 =$$
$$= 0{,}013 + 1{,}075 = 1{,}088 \ см^3 \approx 1{,}1 \ см^3.$$

Поэтому

$$V = \frac{1}{6} \pi d^3 \approx 27{,}4 \ см^3 \pm 1{,}1 \ см^3. \qquad (5)$$

Отсюда предельная относительная погрешность объема

$$\delta_V = \frac{1{,}088 \ см^3}{27{,}4 \ см^3} = 0{,}0397 \approx 4\%.$$

Пример 2. Для определения модуля Юнга E по прогибу стержня прямоугольного сечения применяется формула

$$E = \frac{1}{4} \cdot \frac{l^3 p}{a^3 b s},$$

где l — длина стержня, a и b — измерения поперечного сечения стержня, s — стрела прогиба, p — нагрузка.

Вычислить предельную относительную погрешность при определении модуля Юнга E, если $p = 20$ *кГ*; $\delta_p = 0{,}1\%$; $a = 3$ *мм*; $\delta_a = 1\%$; $b = 44$ *мм*; $\delta_b = 1\%$; $l = 50$ *см*; $\delta_l = 1\%$; $s = 2{,}5$ *см*; $\delta_s = 1\%$.

Решение. $\ln E = 3 \ln l + \ln p - 3 \ln a - \ln b - \ln s - \ln 4$.

Отсюда, заменяя приращения дифференциалами, будем иметь

$$\frac{\Delta E}{E} = 3 \frac{\Delta l}{l} + \frac{\Delta p}{p} - 3 \frac{\Delta a}{a} - \frac{\Delta b}{b} - \frac{\Delta s}{s}.$$

Следовательно,

$$\delta_E = 3\delta_l + \delta_p + 3\delta_a + \delta_b + \delta_s = 3 \cdot 0{,}01 + 0{,}001 + 3 \cdot 0{,}01 +$$
$$+ 0{,}01 + 0{,}01 = 0{,}081.$$

Таким образом, предельная относительная погрешность составляет 0,081, т. е. примерно 8% от измеряемой величины.

Произведя численные расчеты, имеем:

$$E = (2{,}10 \pm 0{,}17) \cdot 10^6 \frac{кГ}{см^2}.$$

§ 17. Обратная задача теории погрешностей

На практике важна также обратная задача: каковы должны быть абсолютные погрешности аргументов функции, чтобы абсолютная погрешность функции не превышала заданной величины.

Эта задача математически неопределенна, так как заданную предельную погрешность Δ_u функции $u=f(x_1, x_2, \ldots, x_n)$ можно обеспечить, устанавливая по-разному предельные абсолютные погрешности Δ_{x_i} ее аргументов.

Простейшее решение обратной задачи дается так называемым *принципом равных влияний*. Согласно этому принципу предполагается, что все частные дифференциалы

$$\frac{\partial f}{\partial x_i}\Delta x_i \quad (i=1, 2, \ldots, n)$$

одинаково влияют на образование общей абсолютной погрешности Δ_u функции $u=f(x_1, x_2, \ldots, x_n)$.

Пусть величина предельной абсолютной погрешности Δ_u задана. Тогда на основании формулы (2) § 16

$$\Delta_u=\sum_{i=1}\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}. \tag{1}$$

Предполагая, что все слагаемые равны между собой, будем иметь

$$\left|\frac{\partial u}{\partial x_1}\right|\Delta_{x_1}=\left|\frac{\partial u}{\partial x_2}\right|\Delta_{x_2}=\ldots=\left|\frac{\partial u}{\partial x_n}\right|\Delta_{x_n}=\frac{\Delta_u}{n}.$$

Отсюда

$$\Delta_{x_i}=\frac{\Delta_u}{n\left|\dfrac{\partial u}{\partial x_i}\right|} \quad (i=1, 2, \ldots, n). \tag{2}$$

Пример 1. Радиус основания цилиндра $R\approx 2$ *м*; высота цилиндра $H\approx 3$ *м*. С какими абсолютными погрешностями нужно определить R и H, чтобы его объем V можно было вычислить с точностью до 0,1 $м^3$?

Решение. Имеем $V=\pi R^2H$ и $\Delta_V=0{,}1$ $м^3$.

Полагая $R=2$ *м*; $H=3$ *м*; $\pi=3{,}14$; приближенно получим:

$$\frac{\partial V}{\partial \pi}=R^2H=12;$$

$$\frac{\partial V}{\partial R}=2\pi RH=37{,}7;$$

$$\frac{\partial V}{\partial H}=\pi R^2=12{,}6.$$

Отсюда, так как $n=3$, то на основании формулы (2) будем иметь:

$$\Delta_\pi = \frac{0{,}1}{3\cdot 12} < 0{,}003;$$

$$\Delta_R = \frac{0{,}1}{3\cdot 37{,}7} < 0{,}001;$$

$$\Delta_H = \frac{0{,}1}{3\cdot 12{,}6} < 0{,}003.$$

Пример 2. Требуется найти значение функции

$$u = 6x^2(\lg x - \sin 2y)$$

с точностью до двух десятичных знаков (после запятой), причем приближенные значения x и y равны соответственно 15,2 и 57°. Найти допустимую абсолютную погрешность этих величин.

Решение. Здесь

$$u = 6x^2(\lg x - \sin 2y) = 6\,(15{,}2)^2(\lg 15{,}2 - \sin 114°) = 371{,}9;$$

$$\frac{\partial u}{\partial x} = 12x(\lg x - \sin 2y) + 6xM = 88{,}54,$$

где $M = 0{,}43429$ — модуль перехода;

$$\frac{\partial u}{\partial y} = -12x^2\cos 2y = +1127{,}7.$$

Для того чтобы результат был верен до двух десятичных знаков, нужно выполнение равенства $\Delta_u = 0{,}005$. Тогда по принципу равных влияний имеем:

$$\Delta_x = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial x}\right|} = \frac{0{,}005}{2\cdot 88{,}54} = 0{,}000028;$$

$$\Delta_y = \frac{\Delta_u}{2\left|\frac{\partial u}{\partial y}\right|} = \frac{0{,}005}{2\cdot 1127{,}7} = 0{,}0000022\ рад = 0'',45.$$

Нередко при решении обратной задачи по принципу равных влияний мы можем столкнуться с таким случаем, когда найденные по формуле (2) предельные абсолютные погрешности отдельных независимых переменных окажутся настолько малыми, что добиться соответствующей точности при измерении этих величин практически невозможно. В таких случаях следует отступить от принципа равных влияний и за счет разумного уменьшения погрешностей одной части переменных добиться увеличения погрешностей другой части переменных.

Пример 3. С какой точностью надо измерить радиус круга $R = 30{,}5$ *см* и со сколькими знаками взять π, чтобы площадь круга была известна с точностью до 0,1%?

Решение. Имеем $s=\pi R^2$ и $\ln s=\ln\pi+2\ln R$. Отсюда

$$\frac{\Delta_s}{s}=\frac{\Delta_\pi}{\pi}+\frac{2\Delta R}{R}=0{,}001.$$

По принципу равных влияний следует положить:

$$\frac{\Delta_\pi}{\pi}=0{,}0005;\quad \frac{2\Delta R}{R}=0{,}0005.$$

Отсюда $\Delta_\pi\leqslant 0{,}0016$ и $\Delta_R\leqslant 0{,}00025R=0{,}0076$ *см*.

Таким образом, следовало бы взять $\pi=3{,}14$ и измерять R с точностью до тысячных долей сантиметра. Ясно, что такая точность измерения практически трудно осуществима. Поэтому выгоднее поступить следующим образом: взять $\pi=3{,}142$; отсюда $\frac{\Delta_\pi}{\pi}=0{,}00013$; тогда $\frac{2\Delta R}{R}=0{,}001-0{,}00013=0{,}00087$ и $\Delta_R\leqslant 0{,}013$ *см*. Такая точность достигается сравнительно легко.

Иногда допускают, что предельная абсолютная погрешность всех аргументов $x_i\,(i=1,\ 2,\ldots,\ n)$ одна и та же. Тогда, полагая

$$\Delta_{x_1}=\Delta_{x_2}=\ldots=\Delta_{x_n},$$

из формулы (1) будем иметь:

$$\Delta_{x_i}=\frac{\Delta_u}{\sum\limits_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|}\qquad (i=1,\ 2,\ldots,\ n).$$

Наконец, можно предположить, что точность измерения всех аргументов x_i $(i=1,\ 2,\ldots,\ n)$ одинакова, т. е. предельные относительные погрешности $\delta_{x_i}(i=1,\ 2,\ldots,\ n)$ аргументов равны между собой:

$$\delta_{x_1}=\delta_{x_2}=\ldots=\delta_{x_n}.$$

Отсюда получим:

$$\frac{\Delta_{x_1}}{|x_1|}=\frac{\Delta_{x_2}}{|x_2|}=\ldots=\frac{\Delta_{x_n}}{|x_n|}=k,$$

где k — общее значение отношений.

Следовательно,

$$\Delta_{x_i}=k\,|x_i|\qquad (i=1,\ 2,\ldots,\ n).$$

Подставляя эти значения в формулу (1), находим:

$$\Delta_u=k\sum_{i=1}^{n}\left|x_i\frac{\partial u}{\partial x_i}\right|$$

и

$$k = \frac{\Delta_u}{\sum_{i=1}^{n} \left| x_i \frac{\partial u}{\partial x_i} \right|}.$$

Таким образом, окончательно имеем:

$$\Delta_{x_i} = \frac{|x_i| \Delta_u}{\sum_{j=1}^{n} \left| x_j \frac{\partial u}{\partial x_j} \right|} \quad (i = 1, 2, \ldots, n).$$

Можно также использовать и другие варианты.

Аналогично решается вторая обратная задача теории погрешности, когда задана предельная относительная погрешность функции и ищутся предельные абсолютные или относительные погрешности аргумента.

Иногда в самой постановке задачи имеются условия, не позволяющие использовать принцип равных влияний.

Пример 4. Стороны прямоугольника $a \approx 5$ *м* и $b \approx 200$ *м*. Какова допустимая предельная абсолютная погрешность при измерении этих сторон, одинаковая для обеих сторон, чтобы площадь S прямоугольника можно было определить с предельной абсолютной погрешностью $\Delta_S = 1$ *м*^{2}?

Решение. Так как

$$S = ab,$$

то

$$\Delta S \approx b \Delta a + a \Delta b$$

и

$$\Delta_S = b\Delta_a + a\Delta_b.$$

Согласно условию задачи

$$\Delta_a = \Delta_b,$$

поэтому

$$\Delta_a = \frac{\Delta_S}{a + b} = \frac{1}{205} \approx 0{,}005 \text{ м} = 5 \text{ мм}.$$

§ 18. Точность определения аргумента для функции, заданной таблицей

В вычислительной практике часто возникает необходимость определить аргумент по значению функции, заданной таблицей. Например, постоянно встречается необходимость определить число по его табличному логарифму или угол по табличному значению какой-либо тригонометрической функции и т. п. Понятно, что погрешность функции вызывает погрешность в определении аргумента.

Пусть имеем таблицу с одним входом для функции $y=f(x)$. Если функция $f(x)$ дифференцируема, то для достаточно малых значений $|\Delta x|$ имеем:

$$|\Delta y| = |f'(x)||\Delta x|.$$

Отсюда

$$|\Delta x| = \frac{|\Delta v|}{|f'(x)|}, \tag{1}$$

или

$$\Delta_x = \frac{1}{|y'|}\Delta_y.$$

Применим формулу (1) к некоторым наиболее распространенным табулированным функциям.

А. Логарифмы

Пусть $y=\ln x$, тогда $y'=\frac{1}{x}$.

Отсюда

$$\Delta_x = x\Delta_y. \tag{2}$$

Если же $y=\lg x$, то $y'=\frac{M}{x}$, где $M=0{,}43429$;

$$\Delta_x = \frac{1}{M}x\Delta_y = 2{,}30\,x\Delta_y. \tag{2'}$$

Отсюда, в частности, получаем $\delta_x = 2{,}30\Delta_y$, т. е. предельная относительная погрешность числа в таблице десятичных логарифмов равна примерно $2\frac{1}{2}$-кратной предельной абсолютной погрешности логарифма этого числа.

Б. Тригонометрические функции

1. Если $y=\sin x\left(0<x<\frac{\pi}{2}\right)$, то $y'=\cos x$ и, следовательно,

$$\Delta_x = \Delta_y \sec x \ \textit{рад}. \tag{3}$$

2. Для функции

$$y=\operatorname{tg} x\left(0<x<\frac{\pi}{2}\right)$$

имеем

$$y'=\sec^2 x$$

и

$$\Delta_x = \Delta_y \cos^2 x \ \textit{рад}. \tag{4}$$

3. Если $y=\lg(\sin x)\left(0<x<\frac{\pi}{2}\right)$, то

$$y'=M\operatorname{ctg} x \text{ и } \Delta_x = 2{,}30\operatorname{tg} x\Delta_y \ \textit{рад}. \tag{5}$$

4. Положим $y=\lg(\operatorname{tg}x)\left(0<x<\frac{\pi}{2}\right)$, тогда

$$y'=\frac{2M}{\sin 2x} \text{ и } \Delta_x=1{,}15\sin 2x\Delta_y \text{ рад}. \tag{6}$$

Так как, очевидно, $\frac{\sin 2x}{2}<\operatorname{tg}x$ при $0<x<\frac{\pi}{2}$, то из формул (5) и (6) следует, что угол x по таблице логарифмов тангенсов определяется точнее, чем по таблице логарифмов синусов.

В. Показательная функция

Если $y=e^x$, то $y'=e^x$ и

$$\Delta_x=\frac{\Delta_y}{e^x} \tag{7}$$

или

$$\Delta_x=\frac{\Delta_y}{y}.$$

Пример 1. С какой точностью можно определить число $x\approx 5000$, пользуясь четырехзначной таблицей десятичных логарифмов?

Решение. По формуле (2′) получаем:

$$\Delta_x=2{,}30\cdot 5000\cdot\frac{1}{2}\cdot 10^{-4}\approx 0{,}6,$$

т. е. число x имеет примерно четыре верные цифры.

Пример 2. Найти погрешность в определении угла $x\approx 60°$:

а) по пятизначной таблице логарифмов синусов,

б) по пятизначной таблице логарифмов тангенсов.

Решение. Для первого случая по формуле (5) имеем:

$$\Delta_x=2{,}30\cdot\sqrt{3}\cdot\frac{1}{2}\cdot 10^{-5} \text{ рад}=0{,}00002 \text{ рад}\approx 4''.$$

Во втором случае по формуле (6) получаем:

$$\Delta_x=1{,}15\cdot\sqrt{3}\cdot\frac{1}{2}\cdot 10^{-5} \text{ рад}\approx 0{,}000005 \text{ рад}\approx 1'',$$

т. е. погрешность в четыре раза меньше.

§ 19. Способ границ

Обычно применяемая оценка погрешности функции (§ 16, формула (2)) является приближенной, так как эта оценка основана на пренебрежении произведениями ошибок. В некоторых случаях требуется иметь точные границы для искомого значения функции если известны границы изменения ее аргументов. Проще всего,

этого можно добиться, используя *способ двойных вычислений*, иначе называемый *способом границ*.

Пусть

$$u = f(x_1, x_2, \ldots, x_n)$$

— непрерывно дифференцируемая функция, монотонная по каждому аргументу x_i $(i = 1, 2, \ldots, n)$. Для этого достаточно предположить, что производные $\frac{\partial f}{\partial x_i}$ $(i = 1, 2, \ldots, n)$ сохраняют постоянный знак в рассматриваемой области ω изменения аргументов. Допустим, что

$$\underline{x_i} < x_i < \overline{x_i} \quad (i = 1, 2, \ldots, n), \tag{1}$$

причем параллелепипед (1) целиком принадлежит области ω.

Положим, что $\tilde{x}_i = \underline{x_i}$, $\hat{x}_i = \overline{x_i}$, если функция f — возрастающая по переменному x_i, и $\tilde{x}_i = \overline{x_i}$, $\hat{x}_i = \underline{x_i}$, если функция f — убывающая по переменному x_i.

Тогда, очевидно,

$$\underline{u} < u < \overline{u}, \tag{2}$$

где

$$\underline{u} = f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$$

и

$$\overline{u} = f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n).$$

Заметим, что переменные $\tilde{x}_i$ $(i = 1, 2, \ldots, n)$ и результат действий f над ними можно округлять лишь в сторону уменьшения величины $\underline{u}$, а переменные $\hat{x}_i$ $(i = 1, 2, \ldots, n)$ и результат действий f над ними можно округлять лишь в сторону увеличения величины $\overline{u}$. При этих обстоятельствах будет гарантировано строгое выполнение неравенства (2). В частном случае, если функция f — монотонно возрастающая по каждому аргументу x_i $(i = 1, 2, \ldots, n)$, то имеем просто

$$f(\underline{x_1}, \underline{x_2}, \ldots, \underline{x_n}) < u < f(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n}). \tag{3}$$

Пример. Алюминиевый цилиндр с диаметром основания $d = 2$ *см* $\pm 0{,}01$ *см* и высотой $h = 11$ *см* $\pm 0{,}02$ *см* весит $p = 93{,}4\,\Gamma \pm \pm 0{,}001\,\Gamma$. Определить удельный вес γ алюминия и оценить его предельную абсолютную погрешность.

Решение. Объем цилиндра равен

$$v = \frac{\pi d^2}{4} h;$$

отсюда

$$\gamma = \frac{p}{v} = \frac{4p}{\pi d^2 h}. \tag{4}$$

Из формулы (4) вытекает, что в области $p>0$, $d>0$, $h>0$ функция γ — возрастающая по аргументу p и убывающая по аргументам d и h. Согласно условию задачи имеем:

$$1{,}99\ см \leqslant d \leqslant 2{,}01\ см;$$
$$10{,}98\ см \leqslant h \leqslant 11{,}02\ см;$$
$$93{,}399\ г \leqslant p \leqslant 93{,}401\ г.$$

Кроме того,

$$3{,}14159 < \pi < 3{,}1416.$$

Поэтому

$$\underline{\gamma} = \frac{4 \cdot 93{,}399}{3{,}1416 \cdot 2{,}01^2 \cdot 11{,}02} = 2{,}671 \frac{г}{см^3}$$

(с недостатком) и

$$\overline{\gamma} = \frac{4 \cdot 93{,}401}{3{,}14159 \cdot 1{,}99^2 \cdot 10{,}98} = 2{,}735 \frac{г}{см^3}$$

(с избытком). Взяв среднее арифметическое, получим:

$$\gamma = 2{,}703 \frac{г}{см^3} \pm 0{,}027 \frac{г}{см^3}, \tag{5}$$

или после округления

$$\gamma = 2{,}70 \frac{г}{см^3} \pm 0{,}03 \frac{г}{см^3}.$$

Для сравнения приведем приближенную оценку погрешности. Используя средние значения аргументов, получим:

$$\gamma = \frac{4 \cdot 93{,}4}{3{,}1416 \cdot 2^2 \cdot 11} = 2{,}703 \frac{г}{см^3}.$$

Логарифмируя формулу (4), имеем:

$$\ln \gamma = \ln 4 + \ln p - \ln \pi - 2 \ln d - \ln h;$$

отсюда, взяв полный дифференциал, получим:

$$\frac{\Delta \gamma}{\gamma} = \frac{\Delta p}{p} - \frac{\Delta \pi}{\pi} - \frac{2\,\Delta d}{d} - \frac{\Delta h}{h}.$$

Следовательно,

$$\delta_\gamma = \delta_p + \delta_\pi + 2\delta_d + \delta_h = \frac{0{,}001}{93{,}4} + \frac{0{,}00001}{3{,}1416} + \frac{2 \cdot 0{,}01}{2} + \frac{0{,}02}{11} =$$
$$= 1{,}07 \cdot 10^{-5} + 3{,}18 \cdot 10^{-6} + 10^{-2} + 1{,}82 \cdot 10^{-3} = 1{,}183 \cdot 10^{-2}.$$

Далее, находим:

$$\Delta_\gamma = \delta_\gamma \cdot \gamma = 1{,}183 \cdot 10^{-2} \cdot 2{,}703 = 3{,}2 \cdot 10^{-2} \frac{г}{см^3}.$$

Таким образом, приближенно имеем:

$$\gamma = 2{,}703 \frac{Г}{см^3} \pm 0{,}032 \frac{Г}{см^3},$$

что очень близко совпадает с точной оценкой (5).

§ 20*. Понятие о вероятностной оценке погрешности

Пусть имеем сумму n слагаемых

$$u = x_1 + x_2 + \ldots + x_n.$$

Тогда предельная абсолютная погрешность суммы, как известно, равна

$$\Delta_u = \Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}. \qquad (1)$$

Отсюда в случае, когда предельные абсолютные погрешности слагаемых одинаковы,

$$\Delta_{x_1} = \Delta_{x_2} = \ldots = \Delta_{x_n} = \Delta,$$

будем иметь:

$$\Delta_u = n\Delta. \qquad (1')$$

Формула (1) дает максимальное возможное значение абсолютной погрешности суммы. Эта предельная погрешность достигается лишь тогда, когда ошибки всех слагаемых: 1) наибольшие из возможных и 2) имеют одинаковые знаки. При большом количестве слагаемых такое неблагоприятное стечение обстоятельств является маловероятным. Фактически ошибки отдельных слагаемых, как правило, имеют различные знаки и, следовательно, частично компенсируют друг друга. Поэтому наряду с теоретической предельной погрешностью суммы Δ_u вводят *практическую предельную погрешность* Δ_u^*, реализуемую с некоторой мерой достоверности.

Ограничимся рассмотрением простейшего случая. Пусть абсолютные погрешности Δx_i ($i = 1, 2, \ldots, n$) слагаемых суммы (1) независимы и подчиняются нормальному закону с одной и той же мерой точности. Положим, что с вероятностью, превышающей число γ, абсолютные погрешности слагаемых не превышают числа Δ, т. е.

$$P(|\Delta x_i| \leqslant \Delta) > \gamma.$$

При этом условии в теории вероятностей доказывается, что с той же мерой достоверности абсолютная погрешность суммы u будет удовлетворять неравенству $|\Delta u| \leqslant \Delta\sqrt{n}$, где n — число слагаемых.

Таким образом, за предельную абсолютную погрешность суммы можно принять число

$$\Delta_u^* = \Delta\sqrt{n}. \qquad (2)$$

Например, складывая 100 чисел с абсолютной погрешностью 0,1, мы получим теоретическую предельную ошибку суммы $\Delta_u = 0{,}1 \cdot 100 = 10$. Фактически же можно ожидать, что эта ошибка не превзойдет величины $0{,}1 \cdot 10 = 1$.

В частности, рассмотрим среднее арифметическое n чисел

$$\xi = \frac{1}{n}(x_1 + x_2 + \ldots + x_n).$$

Согласно строгой теории предельная абсолютная ошибка

$$\Delta_\xi = \frac{1}{n} \cdot n\Delta = \Delta;$$

тогда как с бо́льшей степенью достоверности можно утверждать, что практически

$$\Delta_\xi^* = \frac{\Delta\sqrt{n}}{n} = \frac{\Delta}{\sqrt{n}},$$

т. е. практически достоверно, *что среднее арифметическое приближенных чисел имеет повышенную точность по сравнению с этими числами, причем*

$$\Delta_\xi^* \to 0 \text{ при } n \to \infty.$$

Аналогично для случая умножения n сомножителей с одинаковой относительной предельной погрешностью δ можно доказать, что практическая предельная относительная погрешность произведения определяется формулой

$$\delta_u^* = \delta\sqrt{n}. \tag{3}$$

Литература к первой главе

1. А. Н. Крылов, Лекции о приближенных вычислениях, Изд. 2, АН СССР, Л., 1933, гл. I.
2. Д. А. Вентцель, Е. С. Вентцель, Элементы теории приближенных вычислений, Изд. ВВИА им. Н. Е. Жуковского, М., 1949, гл. I.
3. Дж. Скарборо, Численные методы математического анализа, ГТТИ, 1934, гл. I.
4. Я. С. Безикович, Приближенные вычисления, Гостехиздат, 1949, гл. I и II.
5. Г. М. Фихтенгольц, Математика для инженеров, ГТТИ, 1933, ч. 1, гл. I.
6. В. М. Брадис, Устный и письменный счет. Вспомогательные средства вычислений. Энциклопедия элементарной математики, кн. 1, Гостехиз ат, 1951.