

А. ХУМАЛ

ОТНОСИТЕЛЬНАЯ ЧАСТОТА ПОГРЕШНОСТЕЙ, ОБУСЛОВЛЕННЫХ ОКРУГЛЕНИЕМ В ТАБЛИЦЕ ФУНКЦИИ

Математические таблицы принято составлять так, что при пользовании ими допустима линейная интерполяция. Хотя числа, содержащиеся в таблице, округлены вполне аккуратно, получается путем корректной интерполяции и надлежащего округления иногда не совсем верный результат. Это обстоятельство вообще довольно известно.

Интересными же могут оказаться способы оценки относительной частоты ложных результатов. Если требуется при помощи пятизначной таблицы найти, например, $\log 3,3482$, то нужна та ее страница, где

$$\log 3,348 = 0,52479$$

$$\log 3,349 = 0,52492.$$

Путем простых вычислений ($92 - 79 = 13$ и $13 \cdot 0,2 = 2,6$ или округленно 3, после чего $79 + 3 = 82$) получается вывод, что должно быть $\log 3,3482 = 0,52482$.

В таблицах аргумент берется обычно через равные промежутки (здесь через одну тысячную), называемые шагом таблицы, а значения функции округляются до некоторого определенного порядка (здесь до 10^{-5}). Интерполяция обычно оперирует шагом таблицы, как единицей аргумента, и разностью округленных значений функции, как целым числом (в примере это 13); если разность функции берется так (в единицах порядка таблицы), то вычисляемое произведение (здесь $13 \cdot 0,2$) округляется весьма просто: оно заменяется ближайшим целым числом.

Верность полученного значения 0,52482 легко проверить таблицей более высокого класса точности, например, семизначной таблицей; там $\log 3,3482 = 0,5248114$. Верное пятизначное число, как видно, есть 0,52481, а не 0,52482. Можно было бы надеяться, что такого рода ложные результаты довольно редки, возникая в особо неблагоприятных случаях (как в примере, где правило округления заставляет заменить 2,6 числом 3). Поэтому интересно посмотреть еще, как по тем же данным пятизначной таблицы получается $\log 3,3486$.

Поскольку $13 \cdot 0,6 = 7,8$ или округленно 8, а $79 + 8 = 87$, то выходит, что $\log 3,3486 = 0,52487$. Но и это неверно, ведь по семизначной таблице $\log 3,3486 = 0,5248633$.

Возникновению неверных результатов, должно быть, способствуют округленные значения функции в начале и конце шага (т. е. 0,52479 и 0,52492), использованные при интерполяции. По семизначной таблице $\log 3,348 = 0,5247854$ и $\log 3,349 = 0,5249151$, так что оба пятизначные

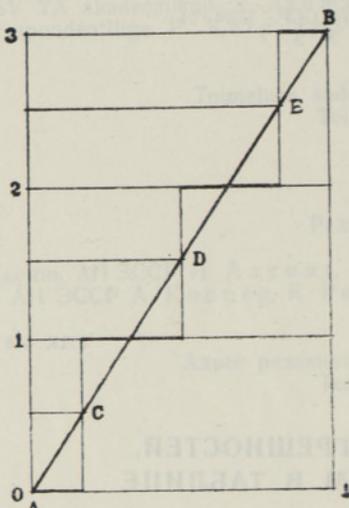


Рис. 1.

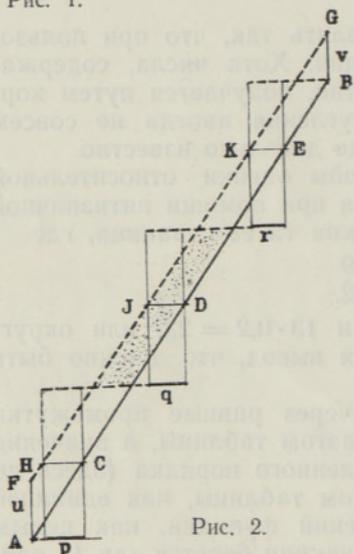


Рис. 2.

округления, хотя вполне корректны, включают заметный избыток. Если полученное интерполяцией слагаемое в силу правила округления тоже появится с избытком, то результат, естественно, может оказаться неверным.

Использование взятых из таблицы округленных значений функции для линейной интерполяции с последующим округлением результата до порядка таблицы на протяжении одного шага наглядно изображено на рис. 1. Представлен случай, когда округленное значение функции в начале шага только на 3 единицы меньше, чем в конце. Линейная интерполяция и округление означают проведение прямой AB через точки $(0; 0)$ и $(1; 3)$ с последующим образованием ступенчатой линии (лесенки) так, что с высоты 0 на высоту 1 происходит вертикальный скачок через точку C прямой AB , имеющую ординату $1/2$, далее скачок с высоты 1 на высоту 2 через точку D , ордината которой $3/2$, и с высоты 2 на высоту 3 через E , где ордината $5/2$.

На рис. 2 та же лесенка сопоставляется с графиком самой функции и получающейся из него лесенкой. Пусть в начале шага функция отличается от своего округленного значения на некоторую величину u , а в конце шага — на величину v (их абсолютное значение не превышает $1/2$). Предположение, что таблица функции допускает линейную интерполяцию, позволяет считать график функции на протяжении шага практически прямолинейным. Эта прямая (при положительных u и v) и соответствующая ей лесенка на рисунке изображены пунктиром. Так, точки H, J и K (ординаты их, соответственно, $1/2, 3/2$ и $5/2$) на прямой FG ,

проведенной через точки $(0; u)$ и $(1; 3 + v)$, практически верно указывают места, где должен происходить вертикальный скачок с одной ступени на другую, чтобы внутри шагового промежутка везде получались безукоризненно верные округления значений функции. При этом выявляются три горизонтальных отрезка (их длины на рисунке обозначены p, q и r), которые можно назвать ошибочными частями ступеней — они состоят из всех тех точек, полученных путем интерполяции и округления, которые оказываются неверными. Вместе с тем ясно, что относительная частота неверных результатов есть $p + q + r$; ведь суммарная длина ступеней одной лесенки равна единице, т. е. длине шага.

Примечательно, что число q выразиимо площадью* трапеции (на

* На рисунках единица абсцисс и единица ординат неравны по длине; при рассмотрении площадей надо учесть, что длина горизонтальных отрезков означает разность абсцисс, а вертикальных — разность ординат.

рисунке пунктированной), имеющей высоту 1 и серединную линию длиной q . А число $p+q+r$ приближенно выражается площадью четырехугольника $ABGF$; в самом деле, хотя верхняя трапеция, имеющая площадь r , оставляет свободным треугольник у вершины G и нижняя тра-

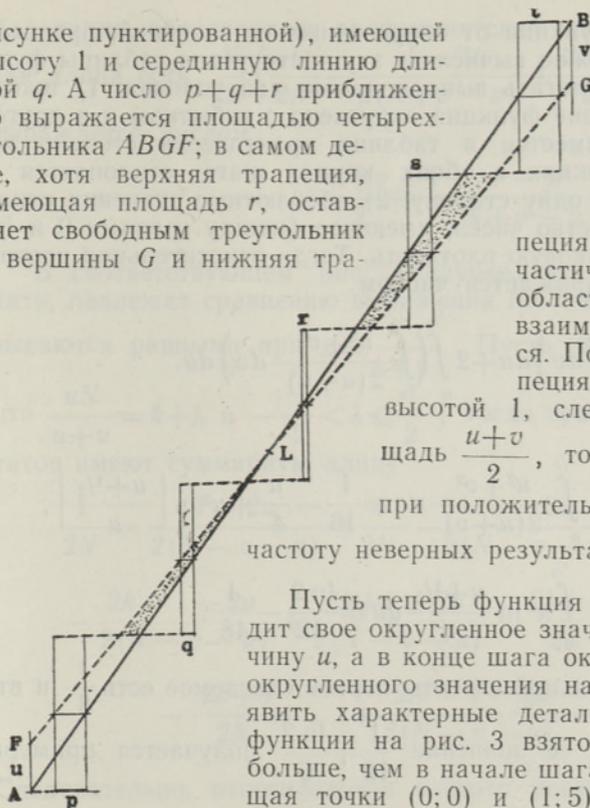


Рис. 3.

пеция, имеющая площадь p , частично выходит за пределы области $ABGF$, эти дефекты взаимно почти компенсируются. Поскольку $ABGF$ есть трапеция с основаниями u , v и высотой 1, следовательно, имеет площадь $\frac{u+v}{2}$, то можно утверждать, что при положительных u , v относительную частоту неверных результатов выражает $\frac{u+v}{2}$.

Пусть теперь функция в начале шага превосходит свое округленное значение на некоторую величину u , а в конце шага оказывается меньше своего округленного значения на величину v . Чтобы выявить характерные детали, округленное значение функции на рис. 3 взято в конце шага на пять больше, чем в начале шага. Прямая AB , соединяющая точки $(0; 0)$ и $(1; 5)$, и соответствующая ей лесенка показывают становление результатов линейной интерполяции и округления. Прямая FG , проведенная через точки $(0; u)$ и $(1; 5-v)$, дает для сравнения лесенку, представляющую верные округленные значения функции всюду внутри шага. Точки, соответствующие неверным результатам, образуют пять отрезков, длины которых обозначены p , q , r , s и t , а относительная частота неверных результатов есть число $p+q+r+s+t$.

Числа q и s оказываются площадями трапеций, меченных на рисунке пунктировкой, число $p+q+r+s+t$ приближенно представимо суммарной площадью треугольников ALF и BLG . Если их основаниями считать u и v , то суммой их высот оказывается 1 (длина шага); вследствие их подобия сами эти высоты равняются, соответственно, числам $\frac{u}{u+v}$ и $\frac{v}{u+v}$, а сумма площадей — числу $\frac{1}{2} \frac{u^2+v^2}{u+v}$.

Таково приближенное выражение относительной частоты погрешностей в случае, соответствующем рис. 3.

Эти оценки частоты могут показаться довольно грубыми. Но следует учесть, что рисунками здесь представлены чрезвычайно простые случаи, когда изменение функции на протяжении шага ограничивается немногими единицами порядка таблицы. Обычно же оно доходит до нескольких десятков, сотен или тысяч единиц; рассматриваемая область (трапеция $ABGF$ или пара треугольников ALF , BLG) тогда состоит практически целиком из таких трапеций, площадь которых правильно выражает длину полученных горизонтальных отрезков.

Относительная частота погрешностей, обусловленных округлением, вычислима индивидуально для каждой таблицы, если известны откло-

нения табулированной функции от округленных ее значений, представленных в таблице. Но можно вычислить эту частоту для таблицы функции вообще, если предполагать выполненными два условия: 1) шагов, где в одном конце значение функции округлено с избытком и в другом конце с недостатком, имеется в таблице практически столько же, сколько таких, где функция в обоих концах шага отклоняется от округленного значения в одну сторону; 2) абсолютные значения отклонений составляют множество чисел, имеющее (между гранями 0 и $1/2$) практически всюду одинаковую плотность. Тогда относительная частота неверных результатов выражается числом

$$2 \int_0^{1/2} \left(\int_0^{1/2} \frac{u+v}{2} dv \right) du + 2 \int_0^{1/2} \left(\int_0^{1/2} \frac{u^2+v^2}{2(u+v)} dv \right) du.$$

Поскольку

$$\int_0^{1/2} \left(\int_0^{1/2} \frac{u+v}{2} dv \right) du = \frac{1}{16}, \quad \int_0^{1/2} \frac{u^2+v^2}{2(u+v)} dv = \frac{1}{16} - \frac{u}{4} + u^2 \ln \left| \frac{u+1/2}{u} \right|,$$

$$\int_0^{1/2} \left(\frac{1}{16} - \frac{u}{4} \right) du = 0, \quad \int_0^{1/2} u^2 \ln \frac{u+1/2}{u} du = \frac{\ln 2}{12} - \frac{1}{48},$$

то в выражении относительной частоты первое слагаемое есть $\frac{1}{8}$ и второе $\frac{\ln 2}{6} - \frac{1}{24}$, а само ее значение $\frac{1}{12} + \frac{\ln 2}{6}$ получается примерно 0,198858.

Таким образом, если таблица функции допускает линейную интерполяцию, то вообще ожидаемо, что из полученных значений функции, округленных до порядка таблицы, около 19,89% окажутся неверными.

Коротко о другом способе оценки частоты погрешностей. Пусть округленное значение функции в конце шага на N единиц порядка таблицы больше, чем в начале шага, и пусть функция превышает свое округленное значение в начале шага на u , а в конце шага на v (тех же единиц). Линейное выражение Nx , которым оперирует интерполяция, имеет значения $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots, N - \frac{1}{2}$ при следующих значениях

аргумента x , соответственно, $\frac{1}{2N}, \frac{3}{2N}, \frac{5}{2N}, \dots, \frac{2N-1}{2N}$. Выражение $u + (N+v-u)x$, дающее верные округленные значения функции внутри шага, равняется тем же числам $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots, N - \frac{1}{2}$ уже при меньших значениях аргумента:

$$\frac{1-2u}{2(N+v-u)}, \frac{3-2u}{2(N+v-u)}, \frac{5-2u}{2(N+v-u)}, \dots, \frac{2N-1-2u}{2(N+v-u)}.$$

Следовательно, те промежутки, где предопределено получение неверных результатов, имеют соответственно длину

$$\frac{1}{2N} - \frac{1-2u}{2(N+v-u)}, \frac{3}{2N} - \frac{3-2u}{2(N+v-u)}, \frac{5}{2N} - \frac{5-2u}{2(N+v-u)}, \dots$$

$$\dots, \frac{2N-1}{2N} - \frac{2N-1-2u}{2(N+v-u)}.$$

Как видно, образовалась арифметическая прогрессия, так что суммарная длина есть $\frac{N}{2} \left(\frac{1}{2N} - \frac{1-2u}{2(N+v-u)} + \frac{2N-1}{2N} - \frac{2N-1-2u}{2(N+v-u)} \right)$ и выражается числом

$$\frac{u+v}{2} + \frac{1}{2} \frac{u^2 - v^2}{N+v-u}.$$

В соответствующем рис. 3 случае, но при N ступенях вместо пяти, подлежат сравнению выражения Nx и $u + (N - v - u)x$; они оказываются равными при $x = \frac{u}{u+v}$. Пусть целое число k берется так, что $\frac{uN}{u+v} = k + \lambda$ и $-\frac{1}{2} < \lambda \leq \frac{1}{2}$; тогда промежутки неверных результатов имеют суммарную длину

$$\begin{aligned} & \frac{1}{2N} - \frac{1-2u}{2(N-v-u)} + \frac{3}{2N} - \frac{3-2u}{2(N-v-u)} + \dots + \frac{2k-1}{2N} - \\ & - \frac{2k-1-2u}{2(N-v-u)} + \frac{2k+1-2u}{2(N-v-u)} - \frac{2k+1}{2N} + \frac{2k+3-2u}{2(N-v-u)} - \\ & - \frac{2k+3}{2N} + \dots + \frac{2N-1-2u}{2(N-v-u)} - \frac{2N-1}{2N}. \end{aligned}$$

Следовательно, относительная частота погрешностей теперь выражается числом

$$\frac{k}{2} \left(\frac{k}{N} - \frac{k-2u}{N-v-u} \right) + \frac{N-k}{2} \left(\frac{N+k-2u}{N-v-u} - \frac{N+k}{N} \right),$$

которое вследствие соотношения $k = \frac{uN}{u+v} - \lambda$ превращается в трехчлен

$$\frac{1}{2} \frac{u^2 + v^2}{u+v} + \frac{1}{2} \frac{u^2 + v^2}{N-u-v} - \frac{\lambda^2}{N} \frac{u+v}{N-u-v}.$$

Что касается остальных случаев, когда Nx сравнивается с $-u + (N - v + u)x$, или же с $-u + (N + v + u)x$, то относительную частоту погрешностей выражают тогда соответственно числа

$$\frac{u+v}{2} + \frac{v^2 - u^2}{2(N+u-v)} \quad \text{и} \quad \frac{1}{2} \frac{u^2 + v^2}{u+v} - \frac{1}{2} \frac{u^2 + v^2}{N+u+v} - \frac{\lambda^2}{N} \frac{u+v}{N+u+v}.$$

Как видно, первым способом оценки относительной частоты погрешностей, не учитывающим зависимости от N , уже получены правильно все те члены, которые не содержат N . Поэтому к выводу об относительной частоте неверных результатов, полученному для таблицы функции вообще, остается только добавить уточняющее выражение, зависящее от N . Ввиду того, что

$$\frac{1}{2} \left(\frac{u^2 - v^2}{N+v-u} + \frac{v^2 - u^2}{N+u-v} \right) + \frac{1}{2} \left(\frac{u^2 + v^2}{N-u-v} - \frac{u^2 + v^2}{N+u+v} \right) -$$

$$\begin{aligned}
 & -\frac{\lambda^2}{N} \left(\frac{u+v}{N-u-v} + \frac{u+v}{N+u+v} \right) = \\
 & = \frac{(u^2-v^2)(u-v)}{N^2-(u-v)^2} + \frac{(u^2+v^2)(u+v)}{N^2-(u+v)^2} - \frac{2\lambda^2(u+v)}{N^2-(u+v)^2},
 \end{aligned}$$

это выражение есть

$$\int_0^{1/2} \left[\int_0^{1/2} \left(\frac{(u^2-v^2)(u-v)}{N^2-(u-v)^2} + \frac{(u^2+v^2)(u+v)}{N^2-(u+v)^2} - 2\lambda^2 \cdot \frac{u+v}{N^2-(u+v)^2} \right) dv \right] du,$$

и оно явно имеет порядок N^{-2} . Естественно, захочется представить его, хотя бы приближенно, в разложении по степеням того же N . Из соотношений

$$\begin{aligned}
 & \frac{(u^2-v^2)(u-v)}{N^2-(u-v)^2} + \frac{(u^2+v^2)(u+v)}{N^2-(u+v)^2} = \\
 & = \frac{(u^2-v^2)(u-v)}{N^2} \left(1 + \frac{(u-v)^2}{N^2} + \frac{(u-v)^4}{N^4} + \dots \right) + \\
 & + \frac{(u^2+v^2)(u+v)}{N^2} \left(1 + \frac{(u+v)^2}{N^2} + \frac{(u+v)^4}{N^4} + \dots \right),
 \end{aligned}$$

$$(u^2-v^2)(u-v) + (u^2+v^2)(u+v) = 2(u^3+v^3),$$

$$2 \int_0^{1/2} \left[\int_0^{1/2} (u^3+v^3) dv \right] du = \frac{1}{32},$$

$$(u^2-v^2)(u-v)^3 + (u^2+v^2)(u+v)^3 = 2(u^5+3u^3v^2+3u^2v^3+v^5),$$

$$2 \int_0^{1/2} \left[\int_0^{1/2} (u^5+3u^3v^2+3u^2v^3+v^5) dv \right] du = \frac{5}{384}$$

вытекает, что в состав уточняющего выражения можно включить

$$\frac{1}{32N^2} + \frac{5}{384N^4}.$$

Поскольку

$$-2\lambda^2 \cdot \frac{u+v}{N^2-(u+v)^2} = -2\lambda^2 \cdot \frac{u+v}{N^2} \left(1 + \frac{(u+v)^2}{N^2} + \frac{(u+v)^4}{N^4} + \dots \right),$$

то включению подлежит еще только

$$-\frac{2}{N^2} \int_0^{1/2} \left[\int_0^{1/2} \lambda^2(u+v) dv \right] du - \frac{2}{N^4} \int_0^{1/2} \left[\int_0^{1/2} \lambda^2(u+v)^3 dv \right] du,$$

если опять ограничиться слагаемыми до порядка N^{-4} . Но необходимые здесь вычисления усложнены тем, что λ зависит от u , v и N несколько неудобным образом:

$$\lambda = \frac{uN}{u+v} - k,$$

где целое число k взято так, что $-\frac{1}{2} < \lambda \leq \frac{1}{2}$. Если же в вычислениях заменить функцию λ^2 ее средним значением, зависящим только от N , задача сильно упрощается.

Как видно, λ зависит от v так, что получает значения от $1/2$ до $-1/2$ однократно в каждом интервале $\frac{2Nu}{2k+1} - u \leq v < \frac{2Nu}{2k-1} - u$ (короче: $w \leq v < W$). Интервал имеет длину $\frac{4Nu}{4k^2-1}$, и среднее λ^2 в нем

$$\frac{4k^2-1}{4Nu} \int_w^W \left(\frac{uN}{u+v} - k \right)^2 dv = 2k^2 - \frac{1}{4} + \left(2k^3 - \frac{k}{2} \right) \ln \frac{2k-1}{2k+1}.$$

Среднее значение целых чисел k , входящих в определение функции λ для области $0 \leq u < 1/2$, $0 \leq v < 1/2$ и данного N , практически равно среднему значению выражения $\frac{uN}{u+v}$ для той же области, стало быть, числу

$$4N \int_0^{1/2} \left(u \int_0^{1/2} \frac{dv}{u+v} \right) du,$$

т. е. просто $\frac{N}{2}$. Таким образом, λ^2 имеет в среднем значение

$$\frac{N^2}{2} - \frac{1}{4} + \left(\frac{N^3}{4} - \frac{N}{4} \right) \ln \left(1 - \frac{2}{N+1} \right),$$

а если вместо логарифма взять его разложение и в итоге ограничиться членами до порядка N^{-3} , то приближением этого значения оказывается $\frac{1}{12} + \frac{1}{15N^2} - \frac{4}{3N^3}$.

Остается констатировать, что

$$\begin{aligned} -\frac{2}{N^2} \left(\frac{1}{12} + \frac{1}{15N^2} \right) \int_0^{1/2} \left[\int_0^{1/2} (u+v) dv \right] du - \frac{2}{N^4} \cdot \frac{1}{12} \int_0^{1/2} \left[\int_0^{1/2} (u+v)^3 dv \right] du = \\ = -\frac{1}{48N^2} - \frac{47}{1920N^4} \end{aligned}$$

и что достаточно полную добавку, учитывающую зависимость относительной частоты от N , выражает

$$\frac{1}{32N^2} + \frac{5}{384N^4} - \frac{1}{48N^2} - \frac{47}{1920N^4},$$

короче $\frac{1}{96N^2} - \frac{11}{960N^4}$ и округленно только $0,01 N^{-2}$.

Это означает, что относительная частота погрешностей, обусловленных округлением, весьма слабо зависит от N ; она практически равна

числу $\frac{1}{12} + \frac{\ln 2}{6}$ для всех таблиц любого класса точности, допускающих линейную интерполяцию.

И наконец два замечания. Одно касается естественного вопроса о возможностях уменьшения частоты погрешностей, обусловленных округлением. Такая возможность создается, например, путем дополнения применяемого в линейной интерполяции выражения Nx слагаемым

$Px \frac{1-x}{2}$, если для расчета множителя P берется значение табулированной функции в середине шага с большей точностью (на два разряда), чем округленные значения функции в самой таблице. Пусть округленное значение функции в начале шага обозначено α , в конце того же шага обозначено β , а более точное значение в середине шага обозначено μ ; тогда $P = 8\mu - 4(\alpha + \beta)$. Поскольку в составе множителя P только 8μ подлежит округлению до порядка таблицы и к тому же

$0 < \frac{x(1-x)}{2} < \frac{1}{8}$ при $0 < x < 1$, то в значениях двучлена

$Nx + Px \frac{1-x}{2}$ влияние округленности числа P значительно меньше, чем

числа N . Поэтому обстоятельство, что $\alpha + Nx + Px \frac{1-x}{2}$ при $x = 1/2$

превращается в μ , означает практически полное уничтожение влияния округленности чисел α и β в середине шага**. Заметно убывает внутри всего шага относительная частота погрешностей, обусловленных округленностью чисел α и β ; как показывает соответствующая оценка, эта частота получается примерно 10,67%. Если же интерполяцию продолжить до третьей степени (добавлением еще слагаемого, например,

$Qx \frac{1-x^2}{6}$) и для расчета множителей (P и Q) брать более точные значения функции в двух подходящих местах внутри шага, то относительную частоту погрешностей, обусловленных округленностью чисел α и β , можно снизить до 6,74%.

Другое замечание касается критериев оптимальности интерполяции. Возможны некоторые ситуации, позволяющие считать наилучшим заменяющее функцию выражение, если его наибольшее отклонение от функции в данной области минимальное; иные ситуации позволяют признать его оптимальным, если минимален интеграл квадрата отклонения. При выборе интерполяции специально для таблицы функции (где длина шагов константна и значения функции округлены до одинакового порядка) естественно требовать, чтобы значение функции и значение заменяющего выражения после их округления до порядка таблицы совпали в как можно большей части шага.

** Переход к интерполяции второй степени дает возможность повысить класс точности таблицы функции при сохранении длины шага, или же удлинить шаг при сохранении класса точности. Если для шага таблицы десятичного логарифма сохранить длину 0,001 (обычная его длина при пятизначной таблице) и предусмотреть интерполяцию второй степени, то подходящим порядком таблицы оказывается 10^{-12} . Вместо приведенных в примере данных можно было бы из такой таблицы получить

$$\begin{aligned} \log 3,348 &= 0,524785449321 & N &= 12969829 & P &= 38734 \\ \log 3,349 &= 0,524915147540 \end{aligned}$$

и описание процесса интерполяции: $(\frac{1-x}{2} P + N) x$.

Сопоставлением графика функции с графиками выражений, отклоняющихся от функции внутри шага всюду меньше, чем на единицу порядка таблицы, сравнением соответствующих этим графикам ступенчатых линий в случае большого числа ступеней, подсчетом длин неверных частей ступеней, отысканием фигур, площадь которых выражает длину этих отрезков, и суммированием площадей можно установить, что для оценки относительной частоты погрешностей, обусловленных заменой функции приближенным выражением, надо интегрировать на протяжении шага абсолютное значение его отклонений от функции.

Итак, среди «хороших» интерполяций для таблицы функции оптимальна та, которая минимизирует интеграл абсолютного значения отклонений.

A. HUMAL

ÜMARDUSEST TINGITUD VIGADE SUHTELINE SAGEDUS INTERPOLATSIOONI KASUTAMISEL FUNKTSIOONITABELITES

Tabuleeritud funktsiooniväärtuste ümardatus mõjub interpolatsiooni teatavasti nii, et osa ümardatud interpolatsioonitulemusi erineb õigestest funktsiooni ümarväärtustest.

Esitatakse kaks võtet seesuguste vigade suhtelise sageduse hindamiseks lineaarinterpolatsiooni juhul.

A. HUMAL

RELATIVE HÄUFIGKEIT DURCH RUNDUNG BEDINGTER FEHLER BEI INTERPOLATION IN FUNKTIONSTAFELN

Rundungen der Tafelwerte wirken sich bei Interpolation in Funktionstafeln bekanntlich so aus, daß ein Teil gerundeter Interpolationsergebnisse von den richtig gerundeten Funktionswerten abweicht.

Für die Abschätzung der relativen Häufigkeit solcher Fehler im Falle linearer Interpolation werden zwei Verfahren vorgestellt.