数值分析

SHU ZHI FEN XI

史万明　杨骅飞　吴裕树　孙　新　编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

数值分析

史万明　杨骅飞
吴裕树　孙　新　编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

图书在版编目（CIP）数据

数值分析/史万明等编著. —北京：北京理工大学出版社，2002.8
ISBN 7－81045－943－0

Ⅰ.数…　Ⅱ.史…　Ⅲ.计算方法　Ⅳ.O241

中国版本图书馆 CIP 数据核字（2002）第 023127 号

出版发行/ 北京理工大学出版社
社　　址/ 北京市海淀区中关村南大街 5 号
邮　　编/ 100081
电　　话/ (010)68914775(办公室)，68912824(发行部)
网　　址/ http：// www.bitpress.com.cn
电子邮箱/ chiefedit@bitpress.com.cn
经　　销/ 全国各地新华书店
印　　刷/ 北京房山先锋印刷厂
装　　订/ 天津武清高村印装厂
开　　本/ 787 毫米×1092 毫米　16 开本
印　　张/ 21.75
字　　数/ 530 千字
版　　次/ 2002 年 8 月第 1 版　　2002 年 8 月第 1 次印刷
印　　数/ 1～4000 册　　责任校对/ 陈玉梅
定　　价/ 28.00 元　　责任印制/ 王　军

前　言

本书是为工科本科生和硕士研究生编写的教材，它是在原来教材的基础上，结合多年教学经验和科研实践修订而成的。本着重概念、重方法、重应用、重能力培养的原则，从构造算法、分析算法、使用算法三方面组织教材内容。在构造算法上，除阐明算法的构造思想、原理外，通过进一步的归纳和整理，我们尽量使同类算法都由某一基本原理或某一基本方法导出，以便读者易于领会和掌握同类算法的共同特征以及同类算法中不同方法之间的相异特征。在分析算法的有关理论推导中，我们力求深入浅出、通俗易懂，并补充少量基础知识，便于阅读和教学。在算法设计与理论分析中，对每种算法均十分关注其应用条件及使用中的问题。每类算法都配以例题与习题，以助理解和练习。

学习本书所需的数学基础是微积分和线性代数，以及常微分方程和偏微分方程的基本概念。读者可针对工科本科生和硕士研究生所要求的内容进行选材，其中也包含一部分适合高水平学生深入理解的内容，可供选学。全书共十一章，约需 70～80 学时，对不同专业，其具体内容和学时数可作适当增减。

本书作者不仅长期从事本门学科的教学，而且具有长期从事科研项目计算的经历，这种实践形成了本书朴素、求实的风格。希望通过本书的介绍，使读者在较短的时间内比较顺利地掌握这些数值方法的要领和基本技巧，为今后从事科学计算打下牢固的基础。

限于水平，书中疏漏和缺陷之处难免，敬请读者批评指正。

编　者

2002 年 3 月

目　录

第一章　数值计算中的误差

§1　计数与数值

1.1　远古的计数

数是一串符号或字母的约定性组合，用以表示某种事物的量或值的多寡程度。因此数是事物的量或值的抽象表示，通常称为数值。数值来源于计数，它由远古的计数产生而逐步形成了它的表示方法。计数频繁地在日常生活中出现，无法想像一个成人还有不会计数的。可是人类确实有过一个时期，既不知道用火，也不知道计数。

远古的计数现在看不到了，引导我们走向古老年代，帮助我们猜破这个谜的，有三条途径：

①研究语言，研究民间的传说和歌谣。在语言里还保存了许多人类不会写字时代的痕迹。

②观察婴孩怎样学说话和计数，就像会重演一下人类计数发展的某些步骤，对于人类怎样掌握计数，可以得到一些启示。

③研究原始民族。在非洲、南美洲中部以及一些岛屿上，还有一些很落后的部落，与我们五千年前甚至一万年前的祖先差不多，在有些地方还保存着原始生活方式。调查了解后，就能帮助我们知道古时候是怎样计数的。通过以上三个来源的信息，就能大概描绘出我们祖先在发明文字以前是如何计数的。

在人类刚刚学会说话和用火的远古时候，他们只知道两个数：一和二。如果要数的东西不止两个，就简单地说“很多”。近代发现，还有整个部落，数到三觉得很困难了。在婴孩的发育过程中，也有一段时间，只懂得什么是“一”，什么是“二”，但是不易数到三。慢慢地，又添上了越来越多的新数，人们学会了数到“五”，又把两个“五”加起来成为一个“十”，大自然赋予人类的“计数器”帮助我们学会了它，这个计数器就是两只手和十个手指。

“五”和“十”这二个数，在计数发展史上起了很大的作用。关于这一点是有许多迹象的。在很多古代民族语言里，前十个数的名称是和手指的名称一样的。在有些现代民族的语言里，也还保存着这个现象的痕迹。例如在现代意大利语里，“le dita”这个字即表示“到十为止的数字”，也表示“手指”。“屈指一算”也说明早先人类的计数是和手指分不开的。最后，现代的十进制计数法证明了“十”这个数字在计数方法的发展中有多么重大的意义。由此看出，人类首先学会了五个五个地计数，然后把两个五合起来十个十个地计数，中国的算盘就证明了这一点。

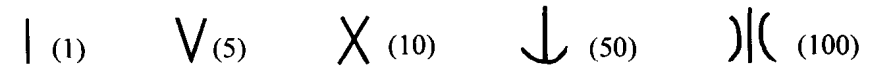
在文字出现前，每一件东西，每一个动作都要用一个特别的符号（一个小小的图画）来表示。开始这些图画都较复杂，经过简化形成为象形文字，这种象形文字至少用了五千年。那时候还没有特别的符号（数字）来表示数，为了改进计数的技巧，必须在两条路里选择一条：或者是转向用简便的文字，即由象形文字改变到用字母来计数；或者是发明一种方法，采用特别的符号来计数。有的民族走了第一条路，如罗马记数法。另一些民族走了第二条路，如巴比伦记数法和中国记数法。

1.2 罗马记数法

字母的发明对于文化的发展有很大的贡献，它也帮助了计数技巧的发展。采用字母来表示数的困难在于，字母是不多的，但是数却有很多。这就是说，不单需要用字母来表示数，而且要发明一种写法，能够用几个字母来写出许多的数。这种用几个字母（或符号）就能写出许多数的方法称做记数法。

罗马记数法特别有意义，因为直到现在，在钟面上、在古老建筑物上都还可以看到它，在书上也还用它来表示章节和世纪等。

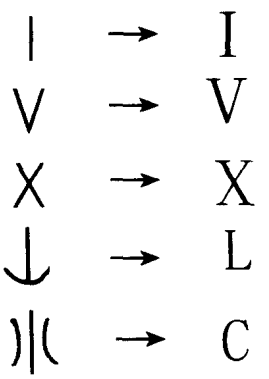
古罗马人在几百年中一直使用着一些奇妙的字母来记数，这些字母的起源直到现在还未搞清楚，它们就是



可以设想，表示一的字母就是采有一个手指的象形文字，表示五的字母就是采用五个手指的象形文字：

而十就是两个五：

但是到了罗马文化最发达的时期（两千年前），这些字母就被和它们相像的拉丁字母代替了。于是改变为：



此外又出现了两个新的字母：D（表示五百）和 M（表示一千），其中 C 和 M 可能是拉丁字母 centum（一百）和 mille（一千）的第一个字母。

罗马人又如何写出各个不同的数呢？要写出数字“二”和“三”，他们就简单地把“一”这个字母重复写二次和三次：Ⅱ，Ⅲ。“四”是这样写的：Ⅳ，在这个写法里，写在五左边的一是要从五里减去的。反之，写在五右边的一是要加到五上面的。因此，六、七、八就写成Ⅵ、Ⅶ、Ⅷ。

再下面就用到 X 这个字母了。“九”写成Ⅸ，接下去是 X、XI、XII、XIII（十、十一、十二、十三）。十四写成 XIV，十五写成 XV 等。二十和三十就写成几个十：XX、XXX。要写四十、五十等就要用到字母 L（五十）。比如四十一写成 XLI，五十、六十、七十就写成 L、LX、LXX。要写九十就使用 C

这个字母，即 X C。一百以内的罗马数字可列写如图 1.1 所示。

I	II	III	IV	V	VI	VII	VIII	IX	X
1	2	3	4	5	6	7	8	9	10
XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
11	12	13	14	15	16	17	18	19	20
XXI	XXII	XXIII	XXIV	XXV	XXVI	XXVII	XXVIII	XXIX	XXX
21	22	23	24	25	26	27	28	29	30
XXXI	XXXII	XXXIII	XXXIV	XXXV	XXXVI	XXXVII	XXXVIII	XXXIX	XL
31	32	33	34	35	36	37	38	39	40
XLI	XLII	XLIII	XLIV	XLV	XLVI	XLVII	XLVIII	IL	L
41	42	43	44	45	46	47	48	49	50
LI	LII	LIII	LIV	LV	LVI	LVII	LVIII	LIX	LX
51	52	53	54	55	56	57	58	59	60
LXI	LXII	LXIII	LXIV	LXV	LXVI	LXVII	LXVIII	LXIX	LXX
61	62	63	64	65	66	67	68	69	70
LXXI	LXXII	LXXIII	LXXIV	LXXV	LXXVI	LXXVII	LXXVIII	LXXIX	LXXX
71	72	73	74	75	76	77	78	79	80
LXXXI	LXXXII	LXXXIII	LXXXIV	LXXXV	LXXXVI	LXXXVII	LXXXVIII	LXXXIX	XC
81	82	83	84	85	86	87	88	89	90
XCI	XCII	XCIII	XCIV	XCV	XCVI	XCVII	XCVIII	IC	C
91	92	93	94	95	96	97	98	99	100

图 1.1

一百后的罗马数字的写法依此类推。102 这个数写成 CII，374 写成 CCCLXXIV等等。从 400 到 899 间的数都要用到 D(五百)，九百写成 CM，一千写成 M。于是 1917 这个数就表为 MCMXVII，1955 表为 MCMLV。用罗马数字写出更大的数也不困难，比如 123849 可写成 CXXIIImDCCCXLIX，其中小写字母“m”表示千，CXXIIIm 表示 123 个千。

罗马字母用来记数还可说是方便的，但是用来计算就不方便了，不论哪种算式的演算，要用罗马数字来做，都是几乎不可能的。

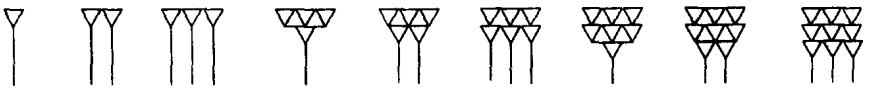
1.3 巴比伦记数法

古代巴比伦人用短棒在柔软的黏土土坯上写字，有的烧成了砖，它们形成了“土坯文件”。学者们在发掘古城时找到了许多这样的文件，其中有的是文契、条约、贸易合同或数学文本。用短棒在柔软的土坯上写字，就使巴比伦人所有的象形文字都是由横的或直的楔形▷或◁构成的，因此就叫做“楔形文字”。通过对数以千计的土坯文件的剖析，比较清楚地了解了古代巴比伦人的生活和文化水平。

大约四千年前，在美索不达米亚平原，就是近东的底格里斯河和幼发拉底河流域，即现在的伊拉克国境内，来到了两个游牧民族：苏美尔人和亚克得人，当时他们都是很文明的民族。过了两个世纪，这两个民族就合并成一个强大的国家——巴比伦。在合并前，两个民族都有自己的质量单位和货币单位，苏美尔人的质量单位叫“明那”，货币单位是“明那银子”。亚克得人的单位比较小，其质量单位为“舍克尔”，合明那的六十分之一。在合并后，上述两种质量单位就同时通用了。随着商业与经济的发展，货币流通量也增加了，巴比伦人也需要更大的单位了。很自然，他们又用比明那大 60 倍的单位作为新的质量单位，因为“六十”这个数在计数中已很习

惯了，这个新单位叫做“塔朗特”；同时也产生了新的货币单位，即一塔朗特银子，它等于 60 明那银子，以上三种单位既是质量单位也是货币单位，每种单位都是较小单位的 60 倍，使得巴比伦人不需要念出和写出比六十更大的数来了。因此，他们只需要使用 59 个符号来记数。

巴比伦人用直立的楔来表示前 9 个数字：



这些符号里的楔排列很合理，念的时候不必去数，因为楔的个数是一眼就可以看出来的。

对于十、二十、三十、四十、五十的符号是用以下宽宽的横写的楔来表示的：

把它们置放在 1～9 个符号的左边就可以获得直至五十九为止的其他记数符号。对于更大的数，就用符号的位置来区分出这些符号是具有什么单位的，它们从右向左的单位分别是舍克尔、明那、塔朗特。比如 2 塔朗特 13 明那 41 个舍克尔就写作：

随着社会的发展，又需要记数技术的改进，必须写出越来越大的数，并且这种数用以表示各种各样的量，从而导致数的本身和它所计量的对象脱开而形成了“抽象数”，即不名数，要写出这些抽象数并不需要想出新的符号，只要运用原有的符号就行了。下面写的数

不再像从前那样表示 21 个明那和 32 个舍克尔，而是表示 32 个初级单位（等于 1）加上 21 个第二级单位（等于 60），这里一个第二级单位有一个初级单位的 60 倍大，更高级的单位依此类推。我们把这种由记号位置不同具有不同单位的记数法叫做进位制记（或计）数法。我们现在还使用着进位制记数法，不同处是我们用的是十进制记数法，而巴比伦人用的是六十进制记数法。

在很长的时期里，巴比伦人没有表示某单位上的符号为零的记号，若出现这种情况，就在该位置上用空位表示。在书写时，这种空位常会大小不一致而产生一些混淆。从某一时候起，在巴比伦的楔形文字中出现了一个新的记号——  （分离符），它相当于零，用来表示一个数里根本不含那一级单位，这样就有

(65)　　(3605)

但巴比伦人却没有想到把这个符号放在数字的末尾，因此在巴比伦人的文化里仍然存在混淆不清的数字，如  表示 1 或 60 或 3600 等。

虽然巴比伦数学家会写出很大的数，但是还不知数是无限多的。另外，这种记数法尚待进一步完善，在保存其进位制的基础上，采用较小的基数来代替基数“60”以达到简化运算规则的目的。也还要学会准确地使用“零”这个符号。这些是由印度人加以改进和完善的。

1.4　印度记数法

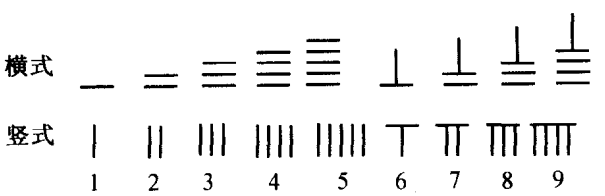
15 世纪以后，印度的科学与艺术得到了繁荣，数学特别被尊重，因为可用它来推算历法、确定四季节气的流转、预测日、月食等。为了写出很大的数，在印度发明一种记数法，它把平常习惯的成十的计数和巴比伦人的进位制记数法结合起来，并且巧妙地使用了“零”这个符号。这种记数法后经阿拉伯人带到了欧洲，排除了其他一切记数法而遍及全世界，它就是我们熟悉的十进制记数法，平常叫做“阿拉伯记数法”，正确地说，应该是“印度记数法”。

十进制记数法由低位至高位（由右向左）依次称为个位、十位、百位、千位、万位……在书写和印刷时，每三位叫做一节，节间要留一个小空位（或打上一个“，”）。在数的念法上，世界各民族有所不同，在欧美，对于大数都有专称，英语中称 100 为 hundred，1 000 为 thousand，1 000 000（百万）为 million，从这以后再大的数各国称法有异，美、法称 1 000 000 000（十亿）为 billion，而英、德称1 000 000 000 000（万亿）为 billion。“百万”这个词在欧洲还是近代产生的，是由 13 世纪到中国来的旅行家马可·波罗想出来的，millione 是意大利文字的百万，它是由两个字合成的，其中 mille 是意大利文的“千”，one 是一个字尾，表示“很多”、“很大”。马可波罗造出这个字来，是为描述“天朝上国”（古时中国的称呼）的无比富庶。

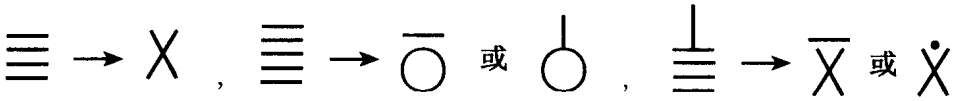
在中国，数字的念法和欧洲各国有所不同，欧洲各国中没有“万”字，“一万”他们叫做十千”，“十万”叫做“一百千”，到了“一千千”就是百万才有新的名称，上面讲过的数的三位分节就是这个原因。而在中国，千上面还有万、十万、百万、千万，直到“万万”才有一个新名，叫做“亿”。再向上就是十亿、百亿、千亿。到了“万亿”称为“兆”。因此按照中国习惯，数的写法按四位一节划分比较方便。

1.5　中国记数法

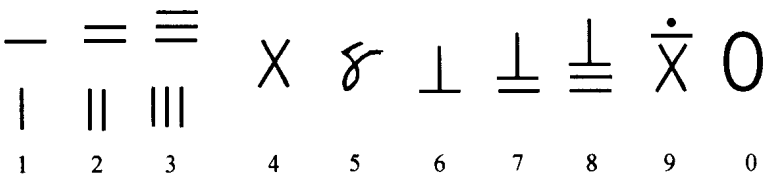
中国古代，为了计算射猎所获得的飞禽和走兽的数目，就用箭来记录，这就是最初的记数法。后来就逐渐用到其他事物的记数，但箭是很长的，用起来不方便，于是就用短竹子削成筹码来代替箭，从这里就产生了象形文字的数字。中国的数字分成横竖两式：



注意到五以上，就用一根筹码来代替五，因为筹码用得太多就看不清了。为何到了“五”才用另一根筹码来代替呢？这当然与我们手上有五个手指分不开的。到了十以上，就不再用新的记号而采用进位制，例如 12 就写成|=或一‖。所以中国古代的记数法，可以说是“用五小进，用十大进”，直到近代也还使用着“一五一十”的口诀。大约到了 6 世纪，人们开始用珠来代替筹码，出现了原始的算盘，到了元代，算盘已在全国风行。从这里看来，中国记数法和巴比伦记数法一样，是采用象形文字和进位制的。书写时采取横竖相间的方式以免混淆不清，例如 1289 就写成|=Ⅲ𝌆。开始中国也没有零，遇到零就空一位。到后来，为简化书写，又把其中几个复杂的数字改变为：



以后𝍠 ⟶ 𝍧，𝍨 或 𝍩 ⟶ 𝍪，而竖式的六、七、八已不再采用，并且造出了 0 这个记号，最后成为近代习惯用的数字：



中国数字的好处，在于容易使人一看便知道各数字所代表的数，另外它也采用十进制。缺点是书写起来不太方便，有的数字要写好几笔。

1.6 通用的记数法

通用的记数法是以印度记数法为原理的 R 进制记数法。其定点形式(称为定点数)为

$$(\alpha_n\alpha_{n-1}\cdots\alpha_1\alpha_0.\,\beta_1\beta_2\cdots\beta_m)_R$$
$$=\alpha_nR^n+\alpha_{n-1}R^{n-1}+\cdots+\alpha_1R^1+\alpha_0R^0+\frac{\beta_1}{R^1}+\frac{\beta_2}{R^2}+\cdots+\frac{\beta_m}{R^m}\qquad(1.1)$$

式中，每位数字 $\alpha_i,\beta_j(i=0,1,2,\cdots,n;j=1,2,\cdots,m)$ 是介于 0 与 $R-1$ 间的正整数。(1.1)式左边是数的书写形式，右边是该数所表达的值，它等于每位数字与其单位乘积之和。该数的总位数称为字长。

通用记数法的浮点形式(称为浮点数)为

$$R^P\cdot(0.\,d_1\,d_2\cdots d_m)_R=R^P\cdot(\frac{d_1}{R^1}+\frac{d_2}{R^2}+\cdots+\frac{d_m}{R^m})\qquad(1.2)$$

式中，P 为阶码，$(0.\,d_1\,d_2\cdots d_m)_R$ 称为尾数，这里 d_i 为介于 0 与 $R-1$ 间的正整数。若 $d_1\neq0$，则称该浮点数为规格化数；否则称为非规格化数。尾数的位数 m 称为字长。在下面三数中：

$$(37.218\ 29)_{10}=10^2\cdot(0.372\ 182\ 9)_{10}=10^3\cdot(0.037\ 218\ 29)_{10}$$

左边是定点数，中间是浮点规格化数，右边是浮点非规格化数。

数的书写形式与其表达的值经常不加区分地统称为数值。采用浮点形式表达数，可以拓宽数值的表达范围，特别是对于甚大(如天文数字)或甚小(如侏儒数字)的定点数可化简其写法。

例如地球的质量是 6 000 000 000 000 000 000 000t 可表为 $10^{22}\cdot(0.6)_{10}$t。氢原子的质量是 0.000 000 000 000 000 000 000 001 65g 可表为 $10^{-23}\cdot(0.165)_{10}$g。

进位制的基数 R 愈大，则数的字长愈短，但其运算规则却随着每位中数字符号的增加而变得更加复杂；反之亦然。

伴随数值的产生和社会发展的需要，又产生了对数值的各种运算及实现运算的种种计算工具，在运算的基础上又产生了各种数值方法，运用它们去解决科学研究或工程技术中的实际问题。特别是电子数字计算机的诞生，是计算数学史甚至人类文明史上的一个里程碑，它使人类获得了高速度、自动化的计算工具，为众多浩繁的数值计算问题的解决展现了光明的前景。同时，也自然地促进数值方法的迅速发展和更新。目前电子数字计算机已成为广泛应用的数值计算工具，但本质上它们能执行的也只是算术运算（加减乘除四则运算）和逻辑运算。而数学问题中的运算范围则是极为广阔的，除算术运算外，还有代数运算、函数运算以及其他种种复杂的运算。数值分析研究的问题，可说是为求解各类数学问题去构造算法、分析算法和使用算法。所谓算法就是由基本运算及规定的运算顺序所构成的完整的解题步骤。构造算法，应以计算机所能执行的运算为依据，尽可能节省存贮量、计算量和提高计算精度。分析算法，就是在数值计算类问题中，主要分析算法的收敛性、稳定性和误差估计等；在逻辑计算类问题中，现在主要研究算法的时间复杂性和空间复杂性，这部分内容已有另一门学科“算法复杂性”专门讲述，它与前一类问题是相互关联的，限于篇幅，本书不多涉及这部分内容。使用算法，就是要注意算法的应用条件、计算过程的控制以及计算机上的使用问题等。

§2 舍入方法与有效数字

数值除来源于计数外，还大量地来源于测量。由于量测工具本身具有的精确度不同，所得到的测量值只能近似地反映出所测物理量的大小。为了衡量其近似的程度，我们引入以下两类误差：绝对误差和相对误差。

2.1 绝对误差与相对误差

设 A 为精确值，a 为近似值，则定义它们之差：

$$\Delta = a - A \tag{1.3}$$

为近似值 a 的绝对误差，简称误差[1]。当 $\Delta>0$ 时，称为正绝对误差；否则称为负绝对误差。由于精确值一般是未知的，因而 Δ 不能求出来。但根据测量误差或计算的情况可以估计出它的上界 ε：

$$|\Delta| = |a - A| < \varepsilon \tag{1.4}$$

则称 ε 为 a 的绝对误差限或误差限。由上式可以推知精确值所界定的范围为

$$a - \varepsilon < A < a + \varepsilon \tag{1.5}$$

有时也用

$$A = a \pm \varepsilon \tag{1.6}$$

来表示。

要刻画近似值的精确程度还必须考虑该精确值本身的大小，以便比较单位数值中所含有

[1] 误差亦可定义为 $\Delta = A - a$，这时误差的符号与本定义相反。

的误差，这就导致引入相对误差的概念。相对误差定义为绝对误差与精确值之比：

$$\delta = \Delta/A \tag{1.7}$$

因 A 一般不知道，实际计算时采用下式

$$\delta = \Delta/a \tag{1.8}$$

来代替，这样代替后，其误差为

$$\frac{\Delta}{A} - \frac{\Delta}{a} = \frac{a - A}{Aa}\Delta = \frac{1}{Aa}\Delta^2$$

当 Δ 很小时，上述误差为 Δ 的高阶无穷小，因此(1.8)式的取法是合理的。相对误差绝对值之上界 η

$$|\Delta/a| < \eta \tag{1.9}$$

称为相对误差限。例如

①$A=0.3\times10^{1}$，$a=0.31\times10^{1}$，则 $\Delta=0.1$，$\delta=0.\dot{3}\times10^{-1}$

②$A=0.3\times10^{-3}$，$a=0.31\times10^{-3}$，则 $\Delta=0.1\times10^{-4}$，$\delta=0.\dot{3}\times10^{-1}$

③$A=0.3\times10^{4}$，$a=0.31\times10^{4}$，则 $\Delta=0.1\times10^{3}$，$\delta=0.\dot{3}\times10^{-1}$

上例表明，计算出的绝对误差虽然差别很大，但它们却有相同的相对误差。显然，对精确性的衡量，只讨论绝对误差是不够的。

绝对误差是有量纲单位的量，而相对误差是一个无量纲的量，有时亦用百分比、千分比等来表示。一般量值范围小的场合以采用绝对误差限为多；量值范围大的场合则采用相对误差限较多。亦有兼有两者的情况，例如在炮兵作业中，采用以下准则来控制对目标的射击命中率：

当 $d_{pm}\leqslant 20\ 000$ m，要求 $|d_{zm}|\leqslant 10$ m

当 $d_{pm}>20\ 000$ m，要求 $|d_{zm}|/|d_{pm}|\leqslant 5/10\ 000$

其中：d_{pm}为炮阵地与目标间的水平距离；d_{zm}为炸点与目标间的水平距离。

2.2 舍入方法

由于我们只能取数值的有限位字长进行计算，因此在计算前，对于一个无限位字长的精确数或字长较长的近似数必须处理成有限位字长的近似数，这种处理方法称为舍入方法。设待舍入处理的数为

$$A = a_0a_1\cdots a_m.\,a_{m+1}a_{m+2}\cdots a_{m+n}a_{m+n+1}\cdots \qquad (a_0 \neq 0) \tag{1.10}$$

今要求对 A 作舍入处理，以获得具有 n 位小数的近似数 a，下面讨论不同的舍入方法。

2.2.1 截断法

我们在 A 中的数字 a_{m+n}与 a_{m+n+1}间将 A 切分为高位部分与低位部分：

$$A = a_0a_1\cdots a_m.\,a_{m+1}\cdots a_{m+n}(\text{高位部分}) + 0.\underbrace{0\cdots0}_{n\text{位}}a_{m+n+1}\cdots(\text{低位部分}) \tag{1.11}$$

截断法就是截取 A 的高位部分作为近似数 a：

$$a = a_0a_1\cdots a_m.\,a_{m+1}\cdots a_{m+n} \tag{1.12}$$

其舍入误差限估计如下：

$$|\Delta| = |a - A| = 0.\underbrace{0\cdots0}_{n\text{位}}a_{m+n+1}\cdots$$

$$\begin{aligned}&\leqslant 0.0\cdots\underbrace{0\,9}_{n\text{位}}\\&\leqslant 0.0\cdots 1=1\times 10^{-n}\end{aligned}\tag{1.13}$$

由上式可见，这种舍入方法导致的舍入误差限不超过近似数 a 最末位数字的一个单位，具有这种精度的近似数 a 称为准确到小数后第 n 位的可靠数，其每一位数字均称为可靠数字。

2.2.2 四舍五入法

此法根据低位部分的最高位数字 a_{m+n+1} 的大小对高位部分的最末位数字 a_{m+n} 进行适当的修改，使 a 的绝对误差限具有最小值，具体方法如下。

(1)四舍情况。

当 $a_{m+n+1}=1,2,3,4$ 时，取近似值 a 为

$$a=a_0a_1\cdots a_m.\,a_{m+1}\cdots a_{m+n}\tag{1.14}$$

其舍入误差限为

$$\begin{aligned}|\Delta|=|a-A|&=0.0\cdots\underbrace{0}_{n\text{位}}a_{m+n+1}\cdots\\&\leqslant 0.0\cdots\dot{4}9\\&\leqslant 0.0\cdots 5=0.5\times 10^{-n}\end{aligned}\tag{1.15}$$

(2)五入情况。

当 $a_{m+n+1}=5,6,7,8,9$ 时，取近似值 a 为

$$a=a_0a_1\cdots a_m.\,a_{m+1}\cdots(a_{m+n}+1)\tag{1.16}$$

其舍入误差限为

$$\begin{aligned}|\Delta|=|a-A|&=|\underbrace{0.0\cdots 1}_{n\text{位}}-\underbrace{0.0\cdots 0}_{n\text{位}}a_{m+n+1}\cdots|\\&\leqslant|0.\ 0\cdots 1-0.0\cdots 05|=0.5\times 10^{-n}\end{aligned}\tag{1.17}$$

综合(1)、(2)可知，用四舍五入法所得的近似值 a，其舍入误差限不超过 0.5×10^{-n}，即其最末位数字的半个单位，与截断法比较，其舍入误差限缩小了一半。

2.2.3 改进的四舍五入法

在四舍五入法中，显然数字 a_{m+n+1} 在五入情况下的个数较四舍情况下的个数多 1，而在大量运算中，数学 1 至 9 在 a_{m+n+1} 位上出现的次数大体上是相同的。因此按四舍五入法对大量运算结果作舍入处理，有可能导致最终结果的数值偏大的弊病。为改进它，可对四舍五入法附加以下补充规定，方法如下。

奇进偶不进法：

$$a_{m+n+1}\begin{cases}<5,\text{将 }a_{m+n}\text{ 后的数字舍去}\\>5,\text{对 }a_{m+n}\text{ 作加 1 处理}\\=5,\begin{cases}\text{当 }a_{m+n}\text{ 为奇数时，对 }a_{m+n}\text{ 作加 1 处理}\\\text{当 }a_{m+n}\text{ 为偶数时，将 }a_{m+n}\text{ 后的数字舍去}\end{cases}\end{cases}\tag{1.18}$$

偶进奇不进法：

$$a_{m+n+1}\begin{cases}<5,\text{将 }a_{m+n}\text{ 后的数字舍去}\\>5,\text{对 }a_{m+n}\text{ 作加 1 处理}\\=5,\begin{cases}\text{当 }a_{m+n}\text{ 为奇数时，将 }a_{m+n}\text{ 后的数字舍去}\\\text{当 }a_{m+n}\text{ 为偶数时，对 }a_{m+n}\text{ 作加 1 处理}\end{cases}\end{cases}\tag{1.19}$$

以上两种方法都能达到进、舍的几率相同，但采用奇进偶不进法更为有利，这是由于作这种舍入处理后的数值均是偶数，一般而论，能把偶数恰好除尽的数要比能把奇数恰好除尽的数要多，这有助于提高计算结果的精度。实践证明，在大量运算中，按上述改进的方法作舍入处理，整个运算过程的舍入误差积累较小。

2.3 有效数字

前面已经证明，通过四舍五入法得到的近似值，其舍入误差限不超过 0.5×10^{-n}，即其最末位数字的半个单位，具有这种精度的近似数 a 称为准确到小数后第 n 位的有效数，其每一位数字均称为有效数字。如果一个数是有效的，则可立即获得该数关于其绝对误差限和相对误差限的估计如下。

推论 1 对于给出的一个有效数，其绝对误差限不大于其最末位数字的半个单位。

推论 2 对于给出的一个有效数，其相对误差限可估计如下：

$$|\delta| = \left|\frac{0.5\times10^{-n}}{a_0\times10^{m}+a_1\times10^{m-1}+\cdots}\right| \leqslant \left|\frac{0.5\times10^{-n}}{a_0\times10^{m}}\right| = \frac{5}{a_0}\times10^{-(m+n+1)} \tag{1.20}$$

因近似数 a 具有 $m+n+1$ 位有效数字，由(1.20)式可见，有效数位愈多，其相对误差就愈小。显见，要想缩小相对误差，最直接而有效的办法就是增加运算中的有效位数。

§3 算术运算中的误差

所谓算术运算指的是+、−、×、÷这四种基本运算，其他的运算都可通过一 定的算法化为一系列算术运算来完成。这里，我们来考虑数据误差在算术运算中的传播规律并对结果的误差进行估计。

设 x^*、y^* 为准确值，x、y 分别为其近似值。则它们的绝对误差分别为 $\Delta x=x-x^*$，$\Delta y=y-y^*$。对于 Δx 和 Δy 常用其主部(指数值的高位部分)近似它们：$\mathrm{d}x\approx\Delta x$，$\mathrm{d}y\approx\Delta y$，它们之间的差别只体现在数值的低位部分，其值甚微可以略去。因此，对于近似值间的算术运算所产生的结果误差的主部可按微分公式来近似估算。

3.1 $c=x\pm y$

$$|\mathrm{d}c| = |\mathrm{d}x\pm\mathrm{d}y| \leqslant |\mathrm{d}x|+|\mathrm{d}y| \leqslant \varepsilon_x+\varepsilon_y \tag{1.21}$$

其中 $|\Delta x|=|x-x^*|\leqslant\varepsilon_x$，$|\Delta y|=|y-y^*|\leqslant\varepsilon_y$

例 1.1 求近似值 285.35，196.87，58.43，4.96 的和，其中每个数的绝对误差限为 0.5×10^{-2}。

解：$285.35+196.87+58.43+4.96=545.61$

和 545.61 的绝对误差限为

$$4\times(0.5\times10^{-2})=0.02<0.5\times10^{-1}$$

因此和值 545.61 应去伪存真作舍入处理成 545.6，它具有 4 位有效数字。