21世纪高等学校教材

袁东锦 编

NUMERICAL ANALYSIS

数值分析

（英文版）

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

本书的出版得到扬州大学出版基金资助

NUMERICAL ANALYSIS

数值分析

（英文版）

袁东锦　编

东南大学出版社
南京

内 容 提 要

本书以英文版的形式介绍各种数值计算方法以及相关的基本概念和理论。内容主要包括误差问题，非线性方程的数值解，插值与多项式逼近，数值积分，解线性方程组的直接法，解大型线性系统的迭代技术，矩阵的特征值和特征向量以及常微分方程初值问题的数值解法等。全书对主要基本算法的推导、构造原理、收敛性、误差估计等进行了较详细的讨论，内容取材适当，由浅入深，各章均有例题和适量的习题以及对各种方法的便于编程的算法描述。

本书可作为理工科院校相关专业"数值分析"课程双语教学或专业英语教学的教材，也可作为非数学专业研究生"数值分析"课程的教材，并可供科技工作者和工程技术人员参考使用。

图书在版编目(CIP)数据

数值分析＝Numerical Analysis/袁东锦编．—南京：东南大学出版社，2005.8

ISBN 7－5641－0091－5

Ⅰ.数… Ⅱ.袁… Ⅲ.数值计算－高等学校－教材－英文 Ⅳ.0241

中国版本图书馆 CIP 数据核字(2005)第 049206 号

东南大学出版社出版发行

(南京四牌楼 2 号 邮编 210096)

出版人：宋增民

江苏省新华书店经销 南京京新印刷厂印刷

开本：700mm×1000mm 1/16 印张：17.5 字数：252 千字

2005 年 8 月第 1 版 2006 年 5 月第 2 次印刷

印数：1501～2500 册 定价：38.00 元

(凡因印装质量问题，可直接向读者服务部调换。电话：025－83792328)

前 言

随着电子计算机的应用日益广泛，科学计算已成为各学科、领域中的一项重要工作。人们在科学研究和生产实际当中会经常碰到数值计算的问题，如何选择与使用适当的数值方法，如何比较、分析数值方法的收敛速度，如何估计、判断计算结果的误差，如何解释数值计算过程中出现的异常现象等等，要解决诸如此类的问题，都需要学习“数值分析”（又称“计算方法”）这一课程。

当今，在各类理工科院校中凡开设了“线性代数”（或“高等代数”）、“数学分析”（或“高等数学”）课程的专业，几乎都要开设此课程，相关专业的研究生更不例外，这是一门内容极其丰富、思想方法深刻而又有着自身理论体系的课程。该课程在国内外都受到相当程度的重视，随着其内容体系的日趋成熟，适用于各种不同层次的中外文版本的教材、专著已比比皆是。既如此，为何又要编写英文版的《数值分析》呢？为的是将高校的教学改革不断推向深入，使我国的高等教育更好地与国际接轨。全国各地高校都在占一定比例的课程上试行双语教学，以使同学们在学习专业知识的过程中借鉴西方先进的东西，了解学科前沿的状况，同时提高使用英语的能力。扬州大学数学科学学院也不例外，从 2002 年开始，我们实施了“对计算方法课程推行双语教学、纯英语教学”的教改课题，这一课题的内容就包括使用自编的英文版教材，在学生中先试行双语教学再逐步向纯英语教学过渡。扬州大学在 2004 年国家教育部组织的本科教学水平评估中以优秀的成绩一举通过，双语教学的教改课题也是其中的亮点之一。那么，为何不使用一本外文原版的教科书呢？笔者在比较了十数本原版教材之后以为，各种原版的教材虽都有其独到的长处，然各教材也不是每章、每节皆完美无缺，正所谓尺有所短，寸有所长。还真难找到一本原版的教材作为我们理想、适合的双语教学的教科书。为此，我们在充分参考和借鉴了多种现行的西方教材（如美国著名的哈佛大学、麻省理工学院和斯坦福大学的现行教材）的基础上，结合我们自己的教学实际，即：进度实际、学生实际、课时实际、专业实际等，博采各家之长，自编一本英文版教材，既介绍西方先进的科学的内

容体系，又切合国内、校内的教学实践。另外，从经济的角度说，一本同篇幅的原版教材，价格在数百元至千元，我们必须结合国情，以人为本，能让中国学生只花少于十分之一的代价读到一本相关内容的教材，以达到相同的教学目的又何乐而不为呢？

然当一权威人士得知编写此书的目的是为对中国学生进行双语教学时则大不以为然，坦率地告知于我：搞教育仍以母语为好。这一下子又使本人堕入云雾之中。仔细想来，此话也不无道理，在用语言表达思想和交流信息的过程中，最流畅也最为准确的仍为母语。但无论如何我们不应忘记进行此项教改课题的宗旨，那就是要在高等教育的方方面面减小与西方的差距，这个“轨”才能较顺利地接上去。客观地说，我们的教学内容与西方的教学内容在培养学生创新能力以及运用所学知识解决实际问题能力方面还存在着相当的差距。虽然，我国学生的读书能力不比西方学生差，就考试分数而言，往往能超过他们，但我们的教学认知过程与实际应用过程常产生较严重的脱节现象，所以我们应该承认在这些方面与西方的差距，要在培养学生学好基础知识的同时更加关注对学生创新思维与创新能力的培养。正是基于此种考虑，笔者在编写此书过程中，从内容的编排，例题与习题的取舍，突出重点等一系列环节上力求克服以上问题并试图缩小一些差距。尽管目前所产生的效果还不太明显，但只要在过程中不断探索、总结、完善、提高，通过一段时间的积淀之后，该项教改举措的成果会逐渐显现出来。

编写此书的过程中得到了扬州大学、扬州大学教务处、扬州大学数学科学学院的指导，在精神上的鼓励与关怀，在经济上的扶助与支持，在此表示衷心的感谢。

由于编者水平与能力所限，教材中肯定存在不少谬误和错漏之处，恳请各位读者和有识之士不吝斧正。

编　者

2005 年 7 月

Contents

1 Preliminaries

This chapter contains a short review of those topics from elementary single variable calculus that will repeatedly be needed in later chapters, together with an introduction to the terminology used in discussing convergence, error analysis, and the machine representation of numbers.

1.1 Review of Calculus

Fundamental to the study of calculus are the concepts of **limit** and **continuity** of a function.

Definition 1.1 Let f be a function defined on a set X of real numbers; f is said to have the **limit** L at x_0, written $\lim_{x \to x_0} f(x) = L$, if, given any real number $\epsilon > 0$, there exists a real number $\delta > 0$ such that $|f(x) - L| < \epsilon$, whenever $x \in X$ and $0 < |x - x_0| < \delta$. (See Figure 1.1)

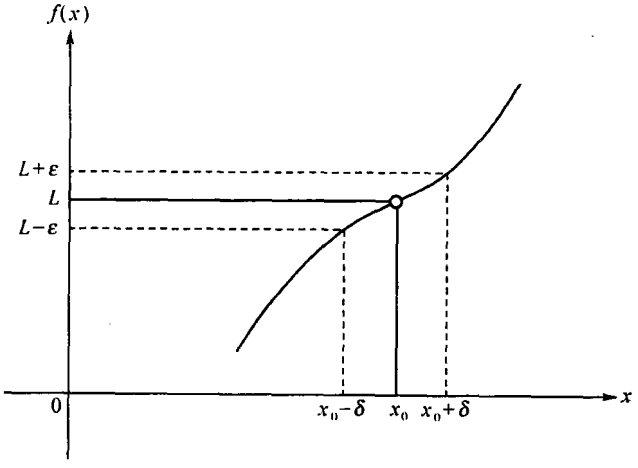

Figure 1.1

Definition 1.2 Let f be a function defined on a set X of real numbers and $x_0 \in X$; f is said to be **continuous** at x_0 if $\lim_{x \to x_0} f(x) = f(x_0)$. The function f is said to be continuous on X if it is continuous at each number in X; $C(X)$ denotes the

set of all functions continuous on X. When X is an interval of the real line, the parentheses in this notation will be omitted. For example, the set of all functions continuous on the closed interval $[a,b]$ will be denoted $C[a,b]$.

In a similar manner, the **limit of a sequence** of real or complex numbers can be defined.

Definition 1.3 Let $\{x_n\}_{n=1}^{\infty}$ be an infinite sequence of real or complex numbers. The sequence is said to **converge** to a number x(called the limit) if, for any $\varepsilon>0$, there exists a positive integer $N(\varepsilon)$ such that $n>N(\varepsilon)$ implies $|x_n-x|<\varepsilon$. The notation $\lim_{n\to\infty} x_n = x$, or $x_n \to x$ as $n\to\infty$, means that the sequence $\{x_n\}_{n=1}^{\infty}$ converges to x.

The following theorem relates the concepts of convergence and continuity.

Theorem 1.4 If f is a function defined on a set X of real numbers and $x_0 \in X$, then the following are equivalent:

(1) f is continuous at x_0;

(2) if $\{x_n\}_{n=1}^{\infty}$ is any sequence in X converging to x_0, then

$$\lim_{n\to\infty} f(x_n) = f(x_0)$$

Definition 1.5 If f is a function defined in an open interval containing x_0, f is said to be **differentiable** at x_0 if

$$\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$$

exists. When this limit exists it is denoted by $f'(x_0)$ and is called the **derivative** of f at x_0. A function that has a derivative at each number in a set X is said to be differentiable on X.

Theorem 1.6 If the function f is differentiable at x_0, then f is continuous at x_0.

The set of all functions that have n continuous derivatives on X is denoted by $C^n(X)$, and the set of functions that have derivatives of all orders at each number in X is denoted by $C^{\infty}(X)$. Polynomial, rational, trigonometric, exponential, and logarithmic functions are in class $C^{\infty}(X)$, where X consists of all numbers at which the functions are defined. When X is an interval of the real line, we will again omit the parentheses in this notation.

The next theorems are of fundamental importance in deriving methods for error estimation. The proofs of these theorems and the other unreferenced results in this section can be found in any elementary calculus text.

Theorem 1.7 (Rolle's Theorem) Suppose $f\in C[a,b]$ and f is differentiable

on (a,b). If $f(a)=f(b)=0$, then a number $c, a<c<b$, exists with $f'(c)=0$. (See Figure 1.2)

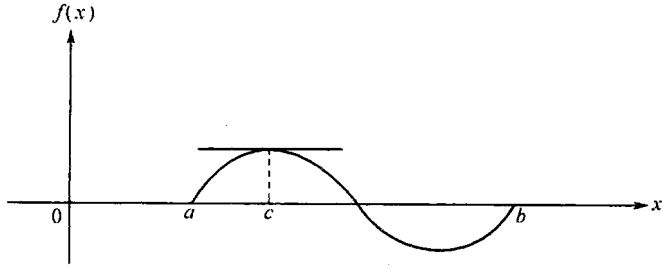

Figure 1.2

Theorem 1.8(Mean Value Theorem) If $f\in[a,b]$ and f is differentiable on (a,b), then a number $c, a<c<b$, exists such that

$$f'(c)=\frac{f(b)-f(a)}{b-a}$$ (See Figure 1.3)

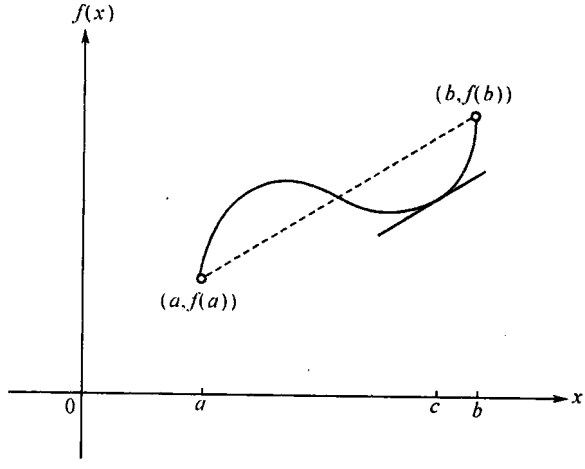

Figure 1.3

Theorem 1.9 (Extreme Value Theorem) If $f\in C[a,b]$, then $c_1, c_2\in[a,b]$ exist with $f(c_1)\leqslant f(x)\leqslant f(c_2)$ for each $x\in[a,b]$. If, in addition, f is differentiable on (a,b), then either $c_i=a$, $c_i=b$, or $f'(c_i)=0$ for each $i=1,2$.

Two other results will be needed in our study of numerical methods. The first is a generalization of the usual Mean Value Theorem for Integrals.

Theorem 1.10(Weighted Mean Value Theorem for Integrals) If $f\in C[a,b]$, g is integrable on $[a,b]$, and $g(x)\geqslant 0$, then there exists a number $c, a<c<b$,

such that

$$\int_a^b f(x)g(x)\mathrm{d}x = f(c)\int_a^b g(x)\mathrm{d}x$$

When $g(x) \equiv 1$, this theorem gives what is called the **average value** of the function over the interval $[a,b]$ (see Figure 1.4). The proof of Theorem 1.10 is not generally given in a basic calculus course, but can be found in any standard advanced calculus text.

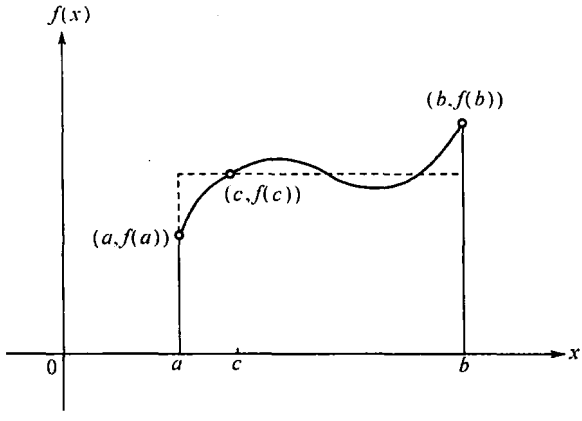

Figure 1.4

The other theorem we will need that is not generally presented in a basic calculus course is derived by applying Rolle's Theorem (Theorem 1.7) successively to $f, f', \cdots$, and finally to $f^{(n-1)}$.

Theorem 1.11 (Generalized Rolle's Theorem) Let $f \in C[a,b]$ be n times differentiable on (a,b). If f vanishes at the $n+1$ distinct numbers $x_0, \cdots, x_n$ in $[a,b]$, then a number c in (a,b) exists with $f^{(n)}(c)=0$.

The next theorem presented is the Intermediate Value Theorem. Although its statement is intuitively clear, the proof is beyond the scope of the usual calculus course. The proof can be found in most advanced calculus texts.

Theorem 1.12 (Intermediate Value Theorem) If $f \in C[a,b]$ and K is any number between $f(a)$ and $f(b)$, then there exists c in (a,b) for which $f(c)=K$ (see Figure 1.5).

Example 1 To show that $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a solution on the interval $[0,1]$, consider the function $f(x) = x^5 - 2x^3 + 3x^2 - 1$. Clearly f is continuous on $[0,1]$ and $f(0)=-1$ while $f(1)=1$. Since $f(0)<0<f(1)$, the intermediate Value Theorem implies that there is a number x, with $0<x<1$, for

which $x^5-2x^3+3x^2-1=0$. □

As seen in Example 1, the Intermediate Value Theorem is important as an aid to determine when solutions to certain problems exist. It does not, however, give a means for finding these solutions. This topic will be discussed more thoroughly in Chapter 2.

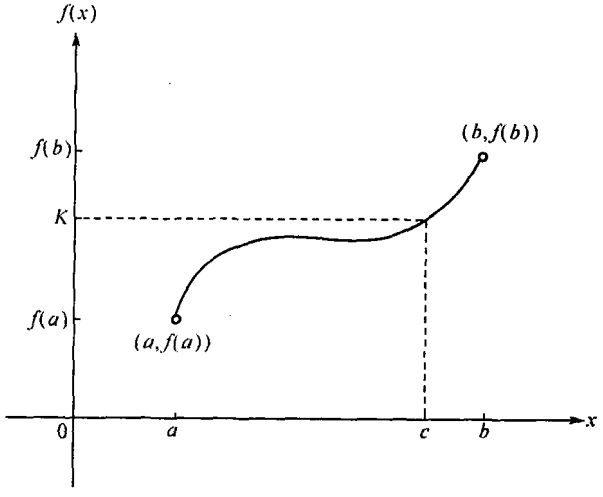

Figure 1.5

The final theorem in this review from calculus describes the development of the Taylor polynomials. The importance of the Taylor polynomials to the study of numerical analysis cannot be overemphasized, and the following result will be used repeatedly.

Theorem 1.13(Taylor's Theorem) Suppose $f\in C^n[a,b]$ and $f^{(n+1)}$ exists on $[a,b]$. Let $x_0\in[a,b]$. For every $x\in[a,b]$, there exists $\xi(x)$ between x_0 and x with

$$f(x)=P_n(x)+R_n(x)$$

where

$$P_n(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

$$=\sum_{k=0}^{n}\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

and

$$R_n(x)=\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}$$

Here $P_n(x)$ is called the **nth-degree Taylor polynomial** for f about x_0 and $R_n(x)$ is called the **remainder term** (or **truncation error**) associated with $P_n(x)$. The infinite series obtained by taking the limit of $P_n(x)$ as $n\to\infty$ is called the **Taylor Series** for f about x_0.

The term **truncation error** generally refers to the error involved in using a truncated or finite summation to approximate the sum of an infinite series. This terminology will be reintroduced in subsequent chapters.

Example 2 Let $f(x)=\cos x$. Since $f\in C^{\infty}(R)$, Theorem 1.13 can be applied for any $n>0$.

For $n=2$ and $x_0=0$, Theorem 1.13 gives

$$\cos x=1-\frac{1}{2}x^2+\frac{1}{6}x^3\sin\xi(x)$$

where $\xi(x)$ is a number between 0 and x.

With $x=.001$, the Taylor polynomial and remainder term is

$$\begin{aligned}\cos .001 &=1-\frac{1}{2}(.001)^2+\frac{1}{6}(.001)^3\sin\xi(x)\\ &=.9999995+(.16\overline{6})\cdot 10^{-9}\sin\xi(x)\end{aligned}$$

where $0<\xi(x)<.001$. (The bar over the last digit in .166 is used to indicate that this digit repeats indefinitely.)

Since $|\sin\xi(x)|<1$, .9999995 can be used as an approximation to cos .001 with assurance of at least nine decimal-place accuracy. Using standard tables, it can be found that

$$\cos .001=.999999500000042$$

so there is actually 13-decimal-place accuracy.

If, in this example, the third-degree Taylor polynomial had been used with $x_0=0$, then

$$\cos x=1-\frac{1}{2}x^2+\frac{1}{24}x^4\cos\xi(x)$$

where $0<\xi(x)<.001$, since $f'''(0)=0$. The approximating polynomial remains the same, and the approximation would still be .9999995, but 13-decimal-place accuracy would be expected since

$$\left|\frac{1}{24}x^4\cos\xi(x)\right|\leqslant\frac{1}{24}(.001)^4(1)\approx 4.2\times 10^{-14}$$

This corresponds more closely to the actual accuracy obtained. □

Exercise

1. Show that the following equations have at least one solution in the given intervals.

(1) $x\cos x-2x^2+3x-1=0$, $[0.2,0.3]$ and $[1.2,1.3]$

(2) $(x-2)^2-\ln x=0$, $[1,2]$ and $[e,4]$

2. Find intervals containing solutions to the following equations.

(1) $x-3^{-x}=0$ (2) $4x^2-e^x=0$

3. Show that $f'(x)$ is 0 at least once in the given intervals.

(1) $f(x)=1-e^x+(e-1)\sin(\pi x/2)$, $[0,1]$

(2) $f(x)=(x-1)\tan x+x\sin\pi x$, $[0,1]$

1.2 Round-Off Errors and Computer Arithmetic

When a calculator or digital computer is used to perform numerical calculations, an unavoidable error, called **round-off error**, must be considered. This error arises because the arithmetic performed in a machine involves numbers with only a finite number of digits, with the result that many calculations are performed with approximate representations of the actual numbers. In a typical computer, only a relatively small subset of the real number system is used for the representation of all real numbers. This subset contains only rational numbers, both positive and negative, and stores a fractional part, called the **mantissa**, together with an exponential part, called the **characteristic**. For example, a single-precision floating-point number used in the IBM 370 or 3000 series consists of a 1-binary-digit sign indicator, a 7-binary-digit exponent with a base of 16, and a 24-binary-digit mantissa. Since 24 binary digits correspond to between 6 and 7 decimal digits, we can assume that this number has at least 6 decimal digits of precision for the floating-point number system. The exponent of 7 binary digit gives a range of 0 to 127, but because of an exponential bias the range is actually -64 to $+63$, that is, 64 is automatically subtracted from the listed exponent.

The **machine number**

0	1000010	10110011000001000000000

precisely represents the decimal number

$$+\left[\left(\frac{1}{2}\right)^1+\left(\frac{1}{2}\right)^3+\left(\frac{1}{2}\right)^4+\left(\frac{1}{2}\right)^7+\left(\frac{1}{2}\right)^8+\left(\frac{1}{2}\right)^{14}\right]\times 16^{66-64}=179.015625$$

since the first binary digit represents the sign, 0 for plus and 1 for minus, the next seven binary digits represent the exponent, and the last twenty-four binary digits represent the mantissa. This machine number is actually used to represent any real number in the interval

$$[179.01561737060546875,\ 179.01563262939453125]$$

since the next smallest machine number is

$$\boxed{0\quad 1000010\quad 101100110000001111111111} = 179.0156097412109375$$

and the next largest machine number is

$$\boxed{0\quad 1000010\quad 101100110000010000000001} = 179.0156402587890625$$

With this representation, the smallest positive number that can be expressed is $16^{-64} \approx 10^{-77}$, and the largest is $16^{63} \approx 10^{76}$. At least one of the four leftmost binary digits for any nonzero number greater than 16^{-64} is required to be one. Consequently, there are 15×2^{28} numbers of the form

$$\pm .d_1 d_2 \cdots d_{24} \times 16^{e_1 e_2 \cdots e_7}$$

which are used by this system to represent all real numbers. Numbers occurring in calculations that have a magnitude of less than 16^{-64} result in what is called **underflow**, and are often set to zero, while numbers greater than 16^{63} result in an **overflow** condition and cause the computations to halt.

The number representation system described above is not standard for all computing machines, but gives an indication of the possible difficulties that can occur. For the remainder of this discussion, we will, for simplicity, assume that machine numbers are represented in the normalized decimal form

$$\pm .d_1 d_2 \cdots d_k \times 10^n, \qquad 1 \leqslant d_1 \leqslant 9, 0 \leqslant d_i \leqslant 9 \tag{1.1}$$

for each $i=2, \cdots, k$, where, from what we have just discussed, the IBM machines have approximately $k=6$ and $-77 \leqslant n \leqslant 76$.

It is useful to consider the representation of an arbitrary real number in the **floating-point** form (1.1). Any positive real number y can be normalized to achieve the form

$$y = .d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n$$

if we assume y is within the numerical range of the machine. The floating-point form (1.1), denoted by $fl(y)$, is obtained by terminating the mantissa of y at k decimal digits. There are two ways of performing this termination. One method is to simply chop off the digits $d_{k+1} d_{k+2} \cdots$ to obtain

$$fl(y) = .d_1 d_2 \cdots d_k \times 10^n$$

This method is quite accurately called **chopping** the number. The other method is to add $5\times10^{n-(k+1)}$ to y and then chop to obtain

$$fl(y)=.\delta_1\delta_2\cdots\delta_k\times10^n$$

The latter method is often referred to as **rounding** the number. In this method if $d_{k+1}\geq5$, we add one to d_k to obtain $fl(y)$; that is, we round up. If $d_{k+1}<5$, we merely chop off all but the first k digits; so we round down. For example, if $k=5$ and rounding is used, we represent π and e as $.31416\times10^1$ and $.27183\times10^1$, respectively. If $k=5$ and chopping is used, we represent π and e as $.31415\times10^1$ and $.27182\times10^1$, respectively.

Since the real numbers with which we are familiar cannot always be represented exactly inside a machine, it is necessary to consider the error due to this finite-digit approximation. The following definition specifies two methods for measuring approximation errors. These methods will be used throughout the text.

Definition 1.14 If p^* is an approximation to p, **absolute error** is given by $|p-p^*|$, and the **relative error** is given by $|p-p^*|/|p|$, provided that $p\neq0$.

Consider the absolute and relative errors in representing p by p^* in the following example.

Example 1

(1) If $p=.3000\times10$ and $p^*=.3100\times10$, the absolute error is .1 and the relative error is $.333\overline{3}\times10^{-1}$.

(2) If $p=.3000\times10^{-3}$ and $p^*=.3100\times10^{-3}$, the absolute error is $.1\times10^{-4}$ and the relative error is $.333\overline{3}\times10^{-1}$.

(3) If $p=.3000\times10^4$ and $p^*=.3100\times10^4$, the absolute error is $.1\times10^3$ and the relative error is $.333\overline{3}\times10^{-1}$.

This example shows that the same relative error, $.333\overline{3}\times10^{-1}$, occurs for widely varying absolute errors. Consequently, as a measure of accuracy, the absolute error may be misleading and the relative error more meaningful. As the following definition indicates, the relative error can be used to tell something about the number of correct digits of an approximation or representation. □

Returning to the machine representation of numbers we see that the floating point representation $fl(y)$ for the number y has the relative error

$$\left|\frac{y-fl(y)}{y}\right|$$

Using k decimal digits in the representation produces an error bound of 10^{-k+1} for chopping and 5×10^{-k} for rounding.