First name: Thomas
Last name: Fuhrmann
Date: 25.11.03
E-Mail: tom.fuhrmann@aon.at
Homework number: 1
Homework Title: Exercise 1.2, 1.22

Problem description (1.2):

What are the absolute and relative errors in approximating π by each of the following quantities?

(a) 3

(b) 3.14

(c) $\frac{22}{7}$

Problem solution:

Find true value of π and select the number of digits in floating point arithmetic.

By definition:

- Absolute error = Approximate value - True value
- Relative error = $\frac{\text{Absolute error}}{\text{True value}}$

Since $\frac{22}{7} = 3.1429$ we determine the error with a precision of 5 digits

Results:

True value for π for 5 digits = 3.1416

(a) Absolute error: $3 - 3.1416 = -0.1416$, Relative error: $\frac{0.1416}{3.1416} = 4.57\%$

(b) Absolute error: $3.14 - 3.1416 = -0.0016$, Relative error: $\frac{0.0016}{3.1416} = 0.051\%$

(b) Absolute error: $3.1429 - 3.1416- = 0.0013$, Rel. error: $\frac{0.0013}{3.1416} = 0.041\%$

Discussion and Comments:

The best of the above approximations is $\frac{22}{7}$.

Problem description (1.22):

Assume that you are solving the quadratic equation $ax^2 + bx + c = 0$, with $a = 1.22$, $b = 3.34$ and $c = 2.28$, using a normalized floating-point system with $\beta = 10$, $p = 3$

(a) What is the computed value of the discriminant $b^2 - 4ac$

(b) What is the correct value of the discriminant in real (exact) arithmetic ?

(c) What is the relative error in the computed value of the discriminant?

Problem solution:

Compute the value of the discriminant with floating-point arithmetic and also determine real arithmetical value.

By definition:

- Absolute error = Approximate value - True value
- Relative error = $\frac{\text{Absolute error}}{\text{True value}}$

A normalized floating point system using $\beta = 10$ and $p = 3$ is defined as follows:

$$x = \pm \left(d_0 + \frac{d_1}{10} + \frac{d_2}{10^2} \right) 10^E$$

Results:

(a) Assuming rounding to nearest: $b^2 - 4ac = (3.34 \cdot 10^0) \cdot (3.34 \cdot 10^0) 4 \cdot 1.22 \cdot 10^0 \cdot 2.28 \cdot 10^0 = 1.12 \cdot 10^1 1.11 \cdot 10^1 = 0.1$

(b) By using 'real' arithmetic: $b^2 - 4ac = 3.34^2 - 4 \cdot 1.22 \cdot 2.28 = 11.1556 - 4 \cdot 1.22 \cdot 2.28 = 0.0292$

(c) Absolute error of equation $= 0.1 - 0.0292 = 0.0708$
Relative error in the computed value $= \frac{0.0708}{0.0292} = 2.424 = 242\%$

Discussion and comments:

By using a normalized floating point system with small precision even small multiplication operations can produce high relative errors.