Suggestions for the maintenance of accuracy of calculations implemented on digital computers

JOHN F. WALSH

Fordham University, Bronx, New York 10458

An approach that selectively uses double precision in calculating statistical quantities is described. Emphasis is placed upon the multiplication and accumulation rather than on the data themselves. In this way, core storage requirements are minimized and good relative error is also achieved.

The purpose of the present note is to illustrate the maintenance of accuracy that can be gained by incorporating two simple procedures of FORTRAN IV as implemented on IBM's 360 series of computers.

A basic feature in a considerable number of computations is the routine of multiplying, either squaring or taking cross-products, and accumulating the individual products. The above can be considered the basic operational unit in the widely used technique of analysis of variance and also in the development of inner products in a variety of matrix manipulations where accuracy is crucial. Analysis of what is involved in these basic routines as implemented on digital computers seems desirable, particularly for behavioral science practitioners. The computational techniques given in popular statistical textbooks for that audience are not well suited for direct FORTRAN coding. Accuracy can be increased by using deviation scores and formulas throughout the computations. This technique requires two passes of the data and is not as easily coded as the raw-score formulas.

The import of the changes to be desired is particularly relevant when raw data must be maintained in core storage for subsequent accessing and the size of this storage is small to modest. Under these circumstances, the option of typing the entire program in double precision using the implicit specification statement is not a realistic alternative that an E could implement to insure accuracy in his computations.

At this point, it is necessary to describe the components that are involved in multiplying and accumulating individual products. Fundamental is the precision, single or double, of three components—the data, the accumulator, and the multiplication. As indicated above, typing the data and instructions as

double precision to insure accuracy is not always feasible. More explicitly, the double precision mode was intended to be used selectively in analyzing experimental data rather than as the sole mode for computations. Consequently, attention must be directed to the two remaining factors—the variable used to accumulate the individual products and the multiplication.

Table 1
Precision of Data, Accumulator, and Multiplication of Four Computing Procedures

Pattern	Data	Accumulator	Multiplication
A	Single	Single	Single
B	Single	Double	Single
C	Single	Double	Double
D	Double	Double	Double

Table 1 summarizes the four patterns of precision among the three components indicated above that are being considered in this note. The iterative computation of the sum of the squares of the first N integers, $1^2 + 2^2 + \dots + N^2$, was used as a baseline in assessing the implications of each of the four patterns on the accuracy of the operations of multiplying and accumulating. Each pattern in Table 1 was utilized in computing an answer for seven values

of N, ranging from 400 to 25,600. The calculations were performed on an IBM Model 50/360 computer.

The definitional formula for the sum of the squares of the first N integers is: $1/6N(N+1)(2N+1)$. This formula was evaluated in double precision for each of the seven values of N and served as the basis for computing the relative error associated with the results obtained iteratively for each of the four patterns in Table 1. Table 2 presents the relative errors for the four procedures that were implemented for each of the seven series of N integers.

Examination of Table 2 indicates that calculation entirely in single precision, Pattern A, degrades very quickly. The accumulation of only 400 products produces very poor relative error. Alternatively, the simple addition of a double precision accumulator in the computation to store the products of the individual single precision terms, Pattern B, is no different from a complete double precision procedure until 6,400 products are summated. Still more striking, Pattern C, which accumulates and multiplies in double precision, is no less accurate than complete double precision for the range of values considered. More importantly, Pattern C illustrates the advantage of the selective use of mixing modes of precision in sensitive calculations.

Implementation of Pattern C is most easily achieved by using the FORTRAN in-line library function DBLE on one of the two single precision terms of the product and adding it to a variable typed as double in an explicit specification statement. The FORTRAN code implementing the principles of Pattern C in the calculation of 1,000 inner products would be:

```
DIMENSION A(1000),B(1000)
DOUBLE PRECISION SUM
SUM=0.D0
DO 10 I=1,1000
10 SUM = SUM+DBLE(A(I))*B(I)
```

Table 2
Relative Errors Associated With the Calculation of the Sum of Squares of N Integers Under Four Patterns of Precision

Number of Integers	Pattern			
	A	B	C	D
400	-.560398E-5	—*	—	—
800	-.888957E-5	—	—	—
1,600	-.534577E-4	—	—	—
3,200	-.169719E-3	—	—	—
6,400	-.263825E-3	-.922635E-7	—	—
12,800	-.330860E-3	-.435739E-7	—	—
25,600	-.105834E-2	-.181438E-6	—	—

*No relative error indicated