D Redell

Default Rules for Rounding
Fixed Precision Floating Point Arithmetic
Ignoring Over/underflow.

The following rules would have to be amended only slightly to allow for over/underflow, which is a nearly independent and much more complicated topic. For simplicity here we consider the "Representable numbers" to be an infinite discrete subset of the continuum of real numbers.

#1 : The representable numbers must include 0, 1 and, if x then $-x$ too.

#2 : Each representable number must be represented uniquely by a symbol string that represents nothing else.

#3 : Any arithmetic operation* which, when executed without roundoff error, would produce a representable number, must actually be executed without error.

#4 : Do not discard information unnecessarily.

#5 : Any arithmetic operation which cannot be executed without roundoff error must result in a representable number nearest what would have been produced in the absence of roundoff error.

#6 : The preceding rule is ambiguous when <u>two</u> representable numbers are nearest the unrounded result. This ambiguity must be resolved in a systematic way which preserves sign symmetry (e.g. $x-y = -(y-x)$) and is "unbiased" in the sense that "drift" cannot occur; e.g. the sequence $x_0, x_1, x_2, \ldots$ defined for arbitrary x_0 and y by $x_{n+1} := (x_n + y) - y$ has $x_1 = x_2 = x_3 = \ldots$

* The arithmetic operations include $+$, $-$, $\times$, $/$, $|\ldots|$, and conversion; and might be extended to include $\sqrt{\ }$ and other FORTRAN functions if the rules above were slightly relaxed.

W. Kahan
Univ. of Calif. @ Berkeley
Sept. 25, 1969