Round-Off Errors

- How big is the round-off errors in a given floating-point number system?
  - ◇ Consider the mantissa only. The rounding results in an *absolute error* bounded by half of the last digit, i.e.,

$$|\epsilon| \leq \frac{1}{2}\beta^{-t}.$$

  - ◇ For any number x that is within the range of the floating-point number system, if we write $x_r = x(1+\delta)$, then $|\delta| \leq \frac{1}{2}\beta^{1-t}$.
- The proof of the above bound on the *relative error* is interesting.
  - ◇ There exists a unique e such that $\beta^{e-1} \leq x < \beta^e$.
  - ◇ In $[\beta^{e-1}, \beta^e)$, numbers are uniformly spaced by β^{e-t}. (Why?)
  - ◇ It follows that $|x_r - x| \leq \frac{1}{2}\beta^{e-t}$.
  - ◇ Hence $\frac{|x_r-x|}{|x|} \leq \frac{\frac{1}{2}\beta^{e-t}}{\beta^{e-1}} = \frac{1}{2}\beta^{1-t}$.
- On an IBM machine ($\beta = 16$), for example, single precision ($t = 6$) gives $|\ delta| \leq 2^{-21} \approx .477 \times 10^{-6}$ whereas double precision ($t = 14$) gives $|\delta| \leq 2^{-53} \approx .111 \times 10^{-15}$.