# Numerical Analysis

Math 370 | MWF 2:30-3:25
Spring 1997 | 201 Fowler

**SUMMARY** (01.17.97) Floating Point Arithmetic and Round-Off Error

**Warm-Up** One of the points which may have been glossed over in the previous class was the notion of a **FLOATING POINT NUMBER**. This is a number which is being used for engineering or scientific purposes and can be written in scientific notation.

$$0.000747 = \qquad\qquad 314.159265 =$$

$$970000000 = \qquad\qquad -42.0 =$$

## Decimal Machine Numbers

We can represent the machine numbers from the previous class as having the form

$$\pm 0.d_1d_2d_3\cdots d_k \times 10^n, \qquad 1 \leq d_1 \leq 9, 0 \leq d_i \leq 9$$

In our specific case $k = 6$ and $-78 \leq n \leq 76$
Any positive real number $y$ can be normalized to be written in the form

$$y = 0.d_1d_2d_3\cdots d_kd_{k+1}\cdots \times 10^n$$

So, how is this number represented by the computer, since it can only use a finite number of digits?

## Chopping

In this case all the digits after $d_k$ are **ignored** ("chopped off")
So,
$fl(y) =$

## Rounding

In this case if the value of $d_{k+1} \geq 5$ then $d_k$ is replaced by $d_k + 1$
So,
$fl(y) =$

## Example

Write down the decimal machine number representation for 3546.16527

**(a)** using chopping

**(b)** using rounding

**Absolute Error and Relative Error**

If $\tilde{p}$ is an approximation to $p$,

the **absolute error** is $|\tilde{p} - p|$, and the **relative error** is $\dfrac{|\tilde{p} - p|}{|p|}$, provided $p \neq 0$

**Example**

Given $p =$ is approximated by $p^* =$

Find the absolute and relative error involved

**Exercise**

Let's compute the relative error involved in chopping a decimal machine number

Let's compute the relative error involved in rounding a decimal machine number