Math 4610 Fundamentals of Computational Mathematics - Floating Point Representation of Numbers.

Any work that is done on a computer boils down to manipulating numbers. A problem with this is that computers have finite resources and the representation of many numbers requires the use of an infinite number of decimal digits. For example, given a circle, the formula for the circumference is

$$C = 2 \times \pi \times r = \pi \times d$$

where r is the radius of the circle and d is the diameter of the circle. The number π is not a rational number. That is, the decimal expansion of this value has an infinite fractional part. The value can be represented as follows:

$$\pi \approx 3.141592653589793...$$

where the ellipsis notation, ..., means the digits never repeat. So, to get an exact representation of π it is necessary to have an infinite number of digits available. Since compouter resources are finite, we must settle for an approximation.
We could use the approximation

$$\pi \approx 3.141592653589793$$

without including an infinite number of digits. One question the should arise is how many digits will provide us with an accurate enough approximation. In some cases, a very crude approximation is enough. In some of our United States, laws have been passed to legally approximate π using a rational number. For example,

$$\pi \approx \frac{22}{7}$$

provides an approximation that will hold up in a courst of law. If you are pouring a circular concrete slab for a water tank it is a good idea to have a concrete estimate for the number π.
Basically, numbers are best represented on a computer using zeros and ones - or in a binary number system. Other common number systems used involve octal or base 8 and hexidecimal or base 16. Another issue that arises in the representation of numbers is numbers that are relatively prime to base 2. As a simple example, consider the representation of the number 1/3 in base 2. The value is

$$\frac{1}{3} = 0.01010101.....$$

where the last pair of digits repeats forever. If a finite number of binary digits are used to represent 1/3, the result is an approximation of the exact value. Note that a base 10 representation of 1/3 is given by the decimal representation

$$\frac{1}{3} = 0.3333333333333.........$$

IEEE standards reference here.....

sign	mantissa	exponent
1 bit	52 bit	11 bit

Content Items:

- **Floating Point Numbers and Roundoff Error Definitions:**
- **Absolue and Relative Error:**
- **Accumulation of Errors in Algorithms:**
- **The General Root Finding Problem in One Variable:**
- **The Intermediate Value Theorem and the Bisection Algorithm:**
- **Stability of the Algorithm:**