

## Numerical Accuracy

# CS 241

## Data Organization using C

Instructor: **Joel Castellanos**  
e-mail: [joel@unm.edu](mailto:joel@unm.edu)  
Web: <http://cs.unm.edu/~joel/>  
Office: Farris Engineering  
Center (FEC) room 321  
Lab Instructor: **Dongye Meng**  
e-mail: [dymeng@cs.unm.edu](mailto:dymeng@cs.unm.edu)



2/8/2009

## What is the Output?

```
#include <stdio.h>
int main()
{ float x = 1.0/3.0;
  float y = 1000.0/3.0;

  printf("x=%15.1f y=%15.1f\n", x,y);
  printf("x=%15.2f y=%15.2f\n", x,y);
  printf("x=%15.3f y=%15.3f\n", x,y);
  printf("x=%15.4f y=%15.4f\n", x,y);
  printf("x=%15.5f y=%15.5f\n", x,y);
  printf("x=%15.6f y=%15.6f\n", x,y);
  printf("x=%15.7f y=%15.7f\n", x,y);
  printf("x=%15.8f y=%15.8f\n", x,y);
  printf("x=%15.9f y=%15.9f\n", x,y);
  printf("x=%15.10f y=%15.10f\n", x,y);
  printf("x=%15.11f y=%15.11f\n", x,y);
}
```

Number of Spaces for  
Right Justified Float

Number of  
Decimal  
Places

2

## Significant Figures or Decimal Places?

```
#include <stdio.h>
int main()
{ float x = 1.0/3.0;
  float y = 1000.0/3.0;
  ...
}
    x=          0.3   y=          333.3
    x=         0.33  y=         333.33
    x=        0.333  y=        333.333
    x=       0.3333  y=       333.3333
    x=      0.33333  y=      333.33334
    x=     0.333333  y=     333.333344
    x=    0.3333333  y=    333.3333435
    x=   0.33333334  y=   333.33334351
    x=  0.333333343  y=  333.333343506
    x= 0.3333333433  y= 333.3333435059
    x= 0.33333334327 y=333.33334350586
```

3

## What is the Output?

```
#include <stdio.h>
int main()
{ float a = 500.0;
  float b = a*a*a;
  float x, y;

  for (x=a; x<a+10.0; x+=1.5)
  { printf("x=%.1f\n", x);
    }

  for (y=b; y<b+10.0; y+=1.5)
  { printf("y=%.1f\n", y);
    }
}
    x=500.0
    x=501.5
    x=503.0
    x=504.5
    x=506.0
    x=507.5
    x=509.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    y=125000000.0
    ...
```

4

## Representation of Floats and Doubles

- In C, the size of a float and of a double are both implementation specific.
- shuttle.cs.unm.edu uses the IEEE 754 standard:

Format	Bytes	Sign	Exponent	Significand (Mantissa)
float	4 (32 bits)	1 bit	7 bits	24 bits (about 7 decimal digits)
double	8 (64 bits)	1 bit	10 bits	53 bits (about 16 decimal digits)

- Note that from float to double, the range increases by only 3 bits but the precision (number of significant digits) more than doubles.

5

## Quiz 2-1: What is the Output?

```
#include <stdio.h>
int main()
{ int i=0;
  float a = 100;
  a = a*a*a*a*a;
  float c = 5;
  float x = 1000000*c + a;
  float z = a;
  for (i=0; i<1000000; i++)
  { z += c;
  }
  x = x/10000;
  z = z/10000;
  printf("x-z=%.1f\n", x-z);
}
```

- a)  $x-z=0.0$
- b)  $x-z=5.0$
- c)  $x-z=-5.0$
- d)  $x-z=500.0$
- e)  $x-z=3.141593$

6

## Quiz 2-2: What is the Output?

```
#include <stdio.h>
int main()
{ int i=0;
  double a = 100;
  a = a*a*a*a*a;
  double c = 5;
  double x = 1000000*c + a;
  double z = a;
  for (i=0; i<1000000; i++)
  { z += c;
  }
  x = x/10000;
  z = z/10000;
  printf("x-z=%.1f\n",x-z);
  7 }
```

- a)  $x-z=0.0$
- b)  $x-z=5.0$
- c)  $x-z=-5.0$
- d)  $x-z=500.0$
- e)  $x-z=3.141593$