Assignment 1

1. What is the absolute error, the relative error and the percent relative error of 2.718281828 as an approximation of *e*?

2. What is the maximum error of a 1st-order Taylor series approximation around $x_0 = 0$ for approximating the value of e^x for a value of $-0.1 < x < 0$?

3. Round the following numbers to 3 significant digits, writing the result in scientific notation:

4852353253.025253

4534.9999

15.8934653

0.00002385

4. Round the following binary numbers to 3 significant bits, writing the result in scientific notation:

110011101010001.1010

1101000.0001

11.111011

0.0001100001

5. The following ten numbers were randomly chosen from a system that produces uniformly distributed digits on an unknown interval [*a*, *b*] of values. What are good estimates of both *a* and *b*?

6.079, 7.235, 5.355, 4.963, 7.182, 5.371, 4.120, 3.393, 6.603, 5.799

6. The minimum and maximum values in Question 5 are 3.393 and 7.235, respectively. Would you describe the technique in Question 5 as more accurate or equally accurate approximations of *a* and *b*?

7. Significant digits are useful, at best as a colloquial but coarse means of describing relative error. Describe, in your own words, why would the word *coarse* would be a good description of the use of significant digits as opposed to just stating the relative error?

8. What value does `-459323` represent using our six-digit representation?

9. What six-decimal-digit representations would be used for 159383, 13.435, and 0.00034125?

10. Complete the following sentences: Adding one more decimal digit to our six-digit representation would decrease the relative error by a factor of ____. Adding one more bit to the double-precision floating-point representation would decrease the relative error by a factor of ____.

11. Add the following pairs of numbers and write the result in the same representation.

```
-459323     +559323

+749133     +705000

+815383     -803999

1 01110110110 1000110100000000000000000000000000000000000000000000
1 01110110101 0001010110100000000000000000000000000000000000000000
```

12. Multiply the follow pairs of numbers and write the result in the same representation

```
-471200     +513200

+492001     +521030

1 10000000000 0010000000000000000000000000000000000000000000000000
1 01111111100 0000000000000000000000000000000000000000000000000000
```

13. The phenomenon of subtractive cancellation says that if two numbers that are almost equal are subtracted, that the result may have less precision than either of the two operands. Why does a similar phenomenon not occur when adding two numbers that have the same sign?

14. Sort the following 10 floating point numbers:

```
a. 0 11001000010 01100010110001010101011001000…0
b. 1 10010011111 11100111101101000101110110010…0
c. 1 10010011101 00100110100001000000100110100…0
d. 1 11000010010 00000000101111111010010001110…0
e. 0 11100111100 00111000010100011110111011100…0
f. 1 00000010111 11011001011100111110010010000…0
g. 0 10101001010 10011110001000001101001010000…0
h. 1 11001100010 00000011101011000011000000110…0
i. 1 01010110011 01011101000001000011000011000…0
j. 1 00100100000 11010001110010000101000010010…0
```

15. Describe the benefit of denormalized numbers by considering what would happen if the following number was divided by four if denormalized numbers did not exist, and what actually happens.

```
0 00000000001 0010000000000000000000000000000000000000000000000000
```

For your reference, this number is 1.125×2^{-1022}.

Acknowledgement: Irene Huang for noting I used the word “course” instead of “coarse.”