Assignment 1

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

$$|e - 2.718281828| = 4.590 \times 10^{-10}$$

$$\frac{|e - 2.718281828|}{|e|} = 1.689 \times 10^{-10}$$

and thus the percent relative error is 1.689×10^{-8}%.

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 $-0.1 < x < 0$?

The 1st-order Taylor series is $e^x \approx 1 + x$ and because the second derivative of e^x is e^x, the error is $\frac{1}{2}e^{\xi}x^2$ where $-0.1 < x \leq \xi \leq 0$. On this range, the maximum value of e^{ξ} is 1 (when $\xi = 0$) and the maximum value of x^2 is when $x = -0.1$, so 0.01, so the error cannot be larger than 0.005.

Aside: you note that the approximation of $e^{-0.1} = 1 - 0.1 = 0.9$, and the actual value is 0.9048374180, and thus, the error is the exact minus the approximation, which is 0.0048374180.

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

4852353253.025253

4534.9999

15.8934653

0.00002385

4.85×10^9, 4.53×10^3, 1.59×10^1, and 2.38×10^{-5}

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

110011101010001.1010

1101000.0001

11.111011

0.0001100001

1.10×2^{14}, 1.11×2^6, 1.00×2^2, and 1.10×2^{-4}

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

Because there are ten digits, it will be

$$\frac{10\min\{\cdots\} - \max\{\cdots\}}{9} = \frac{10 \times 3.393 - 7.235}{9} = 2.9661$$

and

$$\frac{10\max\{\cdots\} - \min\{\cdots\}}{9} = \frac{10 \times 7235 - 3.393}{9} = 7.6618$$

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*?

The estimators in Question 5 of *a* and *b* are better estimators of *a* and *b* as opposed to just taking the minimum and maximum values of the ten numbers, respectively.

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

Note that "3 significant digits" means that the relative error could be any value in the range $(0.5 \times 10^{-3}, 5 \times 10^{-3}]$, so both relative errors of 0.004952 and of 0.0005134 would be described as "3 significant digits." This is a broad range described by a single integer, and thus would be coarse, as opposed to a finer description by the relative error.

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

-9.323×10^{-4}

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

`+541594`, `+501344` and `+453412`

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

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
```

`+559323`, `+749134`, `+814983`, and

```
1 01110110111 0000101111101000000000000000000000000000000000000000
```

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
```

`-493840`, `+522061` and

`0 01111111101 0010000000000000000000000000000000000000000000000000`

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?

Adding two numbers that are close to each other is approximately 2*n*, while subtracting them is approximately zero. In adding two such digits, we will lose one or two bits, while subtracting one from the other will result in something closer to zero.

For example, in our four-digit representation, 3.519 + 3.517 = 7.036 and the first number represents all numbers in the range (3.5185, 3.5195) and the second (3.5165, 3.5175). Thus, the actual answer could be anything in the range (7.035, 7.037), and so in the worst case, the relative error of the answer is 0.0001421.

However, 3.519 – 3.517 = 0.002 and the first number represents all numbers in the range (3.5185, 3.5195) and the second (3.5165, 3.5175). Thus, the actual answer could be anything in the range (0.001, 0.003), and so in the worst case, the relative error of the answer is 100%.

14. Sort the following 10 floating point numbers:

```
a. 0 11001000010 0110001011000101010101011001000…0
b. 1 10010011111 1110011110110100010111011001 0…0
c. 1 10010011101 0010011010000100000010011010 0…0
d. 1 11000010010 0000000010111111101001000111 0…0
e. 0 11100111100 0011100001010001111011101110 0…0
f. 1 00000010111 1101100101110011111001001000 0…0
g. 0 10101001010 1001111000100000110100101000 0…0
h. 1 11001100010 0000001110101100001100000011 0…0
i. 1 01010110011 0101110100000100001100001100 0…0
j. 1 00100100000 1101000111001000010100001001 0…0
```

From the most negative to the most positive:

```
1 11001100010 00000011101011000011000000110…0
1 11000010010 00000000101111111010010001110…0
1 10010011111 11100111101101000101110110010…0
1 10010011101 00100110100001000000100110100…0
1 01010110011 01011101000001000011000011000…0
1 00100100000 11010001110010000101000010010…0
1 00000010111 11011001011100111110010010000…0
0 10101001010 10011110001000001101001010000…0
0 11001000010 01100010110001010101011001000…0
0 11100111100 00111000010100011110111011100…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}.

Without denormalized numbers, this divided by 4 equals zero, while with denormalized numbers, this ends up being

```
0 00000000000 0100100000000000000000000000000000000000000000000000
```

which represents $0.01001_2 \times 2^{-1022} = 1.001_2 \times 2^{-1024} = 1.125 \times 2^{-1024}$. Note that the denormalized number does not have a leading "1." and instead, only the mantissa is stored.

Acknowledgement: Irene Huang for noting I used the word "course" instead of "coarse", Sarosh Zaidi who noticed the product of two negative numbers was still negative, Andrew Kim for noting that the point at which the error is a maximum was wrong.