Assignment I

Question 1

2.

$$\begin{aligned}\frac{|p^*-p|}{|p|} \le 10^{-4} &\Rightarrow |p^*-p| \le |p| * 10^{-4}\\ &\Rightarrow -p(0.0001) \le p^*-p \le p(0.0001)\\ &\Rightarrow p(0.9999) \le p^* \le p(1.0001)\end{aligned} \tag{1}$$

	value	Lower Bound	Upper Bound
π	3.14159265	3.141278491	3.141906809
$\sqrt{2}$	1.41421356	1.414072139	1.414354981

6. $\pi = 3.14159265$, $e = 2.71828182$.

	Rounded Value	Exact Value
$133 + 0.921$	133.9	133.921
$133 - 0.499$	132.5	132.501
$(121 - 0.327) - 119$	1.7	1.673
$(121 - 119) - 0.327$	1.673	1.673
$\frac{\frac{13}{14}-\frac{6}{7}}{2e-5.4}$	1.986	1.953541043
$-10\pi + 6e - \frac{3}{62}$	-15.16	-15.154622677
$(\frac{2}{9}).(\frac{9}{7})$	0.2857	0.285714286
$\frac{\pi-\frac{22}{7}}{\frac{1}{17}}$	-0.017	-0.021496379

10.

```
#include <iostream.h>
#include <math.h>
int main(){
    int f=1;
    double sum=1.0;
    for (int i=1; i<=5; i++) {
        f*=i;
        sum+=(1.0/f);
    }
    cout << "Abs Err: "<< fabs(exp(1)-sum) << " Rel Err:" << fabs((exp(1)-sum)/exp(1)) <<endl;
    return 0;
}

Abs Err: 0.00161516 Rel Err:0.000594185
```

12.

a. 2.

b. 2.05.

c. After replacing the exponential function with its third Maclaurin polynomial we can derive two answers:

- First simplify the equation, then find $f(0.1)$ using the new function which is more accurate. In this case the result is 2.00.

$$f(x) = \frac{e^x - e^{-x}}{x} = \frac{(1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}) - (1-\frac{x}{1!}+\frac{x^2}{2!}-\frac{x^3}{3!})}{x} = 2 + \frac{x^2}{3} \tag{2}$$

- Find e^x and e^{-x} by using their third Maclaurin polynomials. Then use the values in the main formula. In this case the result is 2.05.

Question 2

a. Exponent bias is 7.

b. $(1.00000)_2 * 2^{-6}$ and $(1.11111)_2 * 2^7$ are the smallest and largest no-negative normalized floating point numbers.

c. 2^{-5}

d. $\frac{1}{10} = (0.0\overline{0011})_2$ or any number which has an infinite representation could not fit in this system. The two closest floating point numbers to this number are $(1.10011)_2 * 2^{-4}$ and $(1.10100)_2 * 2^{-4}$

e. $x = (11.011011_2) = (1.10110|1)_2 * 2$, x_- and x_+ are as follows: $x_- = (1.10110)_2 * 2$, $x_+ = (1.10111)_2 * 2$. To round x by using "round to nearest mode" there is a tie, so we choose the one with least significant bit equal to zero which is x_-.
$y = -(11.011011)_2 = -(1.10110|1)_2 * 2$, y_+ and y_- are as follows: $y_+ = -(1.10110)_2 * 2$, $y_- = -(1.10111)_2 * 2$. To round y by using "round to nearest mode" there is a tie, so we choose the one with least significant bit equal to zero which is y_+.