For this document I am implementing the notation: $\{a\}^n$ to mean what is in a, n times in a row. Example $\{101\}^2 = 101101$

Exercise : 5.1
Wright the following numbers in IEEE double form.

(a) 1.5, round up
$1.5_{10} = 1.1_2$
s=0
m=$1\{0\}^{51}$
e=$0\{1\}^{10}$
Thus in double form our number would be: $00\{1\}^{10}1\{0\}^{51}$

(b) 5.1, round to nearest
$5_{10} = 101_2$
There is a simple way to convert decimals to base 2, simply wright the decimal in base ten and double everything after the decimal place. If you continue doing this the ones and zeros before the decimal are the binary rep. This works basically by removing the 2^{-1} term then the 2^{-2} term and so on.
0.1
0.2
0.4
0.8
1.6
1.2
0.4
0.8
1.6
1.2
0.4
We see now that $0.1_{10} = (0.0\{0011\}^\infty)_2$. So $5.1_{10} = (101.0\{0011\}^\infty)_2 = (1.010\{0011\}^\infty)_2 * 2^2$
s=0
m=$010\{0011\}^\infty$
e=$1\{0\}^9 1$
However we need to round the mantissa. Separating the first 52 bits from the rest we get m=$010\{0011\}^{12}0, 011\{0011\}^\infty$. We now round to nearest to get m=$010\{0011\}^{12}0$.
Thus in double form our number would be: $01\{0\}^9 1010\{0011\}^{12}0$.

(c) -5.1, round towards 0.
In exact terms this is the negation of the previous question:
s=1
m=$010\{0011\}^\infty$
e=$1\{0\}^9 1$
However we need to round the mantissa. Separating the first 52 bits from the rest we

get m=$010\{0011\}^{12}0, 011\{0011\}^{\infty}$. We now round to 0 to get m=$010\{0011\}^{12}0$.
Thus in double form our number would be: $11\{0\}^{9}1010\{0011\}^{12}0$.

(d) -5.1, round down.
In exact terms this is the same as the previous question:
s=1
m=$010\{0011\}^{\infty}$
e=$1\{0\}^{9}1$
However we need to round the mantissa. Separating the first 52 bits from the rest we get m=$010\{0011\}^{12}0, 011\{0011\}^{\infty}$. We now round down to get m=$010\{0011\}^{12}1$.
Thus in double form our number would be: $11\{0\}^{9}1010\{0011\}^{12}1$.

Exercise : 5.2
Wright the following numbers in IEEE double form. 50.2

$50_{10} = 25 * 2^1 + 0 * 2^0 = 12 * 2^2 + 1 * 2^1 + 0 * 2^0 = 6 * 2^3 + 0 * 2^2 + 1 * 2^1 + 0 * 2^0 = 3*2^4+0*2^3+0*2^2+1*2^1+0*2^0 = 1*2^5+1*2^4+0*2^3+0*2^2+1*2^1+0*2^0 = 110010_2$.
Converting .2 to binary I get:
0.2
0.4
0.8
1.6
1.2
0.4
0.8
1.6
1.2
0.4
So $.2_{10} = (0.\{0011\}^{\infty})_2$. Now we see that $50.2 = (110010.\{0011\}^{\infty})_2 = (1.10010\{0011\}^{\infty})_2 * 2^5$
s=0
m=$10010\{0011\}^{\infty}$
e=$1\{0\}^{7}100$
However we need to round the mantissa. Separating the first 52 bits from the rest we get m=$10010\{0011\}^{11}001, 1\{0011\}^{\infty}$. We now round to nearest to get m=$10010\{0011\}^{11}010$.
Thus in double form our number would be: $01\{0\}^{7}10010010\{0011\}^{11}010$.

Exercise : 5.3
What is the gap between 2 and the next largest double precision number?

Noting that 2 is
s=0
m=$\{0\}^{52}$
e=1_{10}
We see that the next largest number would be:

s=0
m=$\{0\}^{51}1$
e=1_{10}
Subtracting the two we get $(1.\{0\}^{51}1 - 1)_2 * 2^1 = (0.\{0\}^{51}1)_2 * 2 = 2^{-52} * 2 = 2^{-51}$

Exercise : 5.4
What is the gap between 201 and the next largest double precision number?

Noting that 201 is
s=0
m=$1001001\{0\}^{45}$
e=7_{10}
We see that the next largest number would be:
s=0
m=$1001001\{0\}^{44}1$
e=7_{10}
Subtracting the two we get $(1.1001001\{0\}^{44}1 - 1.1001001\{0\}^{45})_2 * 2^7 = (0.\{0\}^{51}1)_2 * 2^7 = 2^{-52} * 2^7 = 2^{-45}$

Exercise : 5.5
How many normalized double precision numbers are there?

We can solve this using combinations. There are two possible signs. There are 2^{52} possible mantissas. There are $2^{11} - 2$ possible exponents for normalized numbers as two possibilities $\{0\}^{11}$ and $\{1\}^{11}$ don't represent normalized numbers. So there are $2 * 2^{52} * (2^{11} - 2) = 2^{64} - 2^{54}$ normalized double precision numbers.

Exercise : 5.8

In the described system machine ϵ would be $1.001 - 1.000 = .001 = 2^{-3} = 0.125$. The largest number would be $1.111 * 2^1 = (10 - .001) * 10 = 100 - .01 = 4 - 1/4 = 3\frac{3}{4} = 3.75$. The smallest number would be $1.000 * 2^{-1} = .1 = \frac{1}{2} = 0.5$.

Exercise : 5.9

(a) It is true that $a \oplus b = b \oplus a$ since $a \oplus b$ =round$(a + b)$ =round$(b + a) = b \oplus a$.

(b) It is not true that $(a \oplus b) \oplus c = a \oplus (b \oplus c)$. As a example consider the case where $a = 1$, $b = 1/2\epsilon$, $c = 1/2\epsilon$. Where ϵ is the machine precision. Note that $(a \oplus b) \oplus c = (1 \oplus 1/2\epsilon) \oplus 1/2\epsilon = (1) \oplus 1/2\epsilon = 1$. Where as $a \oplus (b \oplus c) = 1 \oplus (1/2\epsilon \oplus 1/2\epsilon) = 1 + \epsilon$.

(c) $(a \otimes b) \oslash c = (ab(1 + \delta_1)) \oslash c = (ab(1 + \delta_1))/c(1 + \delta_2)$ where $\delta_1, \delta_2 \in [-\epsilon/2, \epsilon/2]$. So our relative error would be $|(a \otimes b) \oslash c - ab/c|/|ab/c| = |(ab)/c(1+\delta_2)(1+\delta_1) - ab/c|/|ab/c| =$

$|(1+\delta_2)(1+\delta_1)-1| = |\delta_2+\delta_1+\delta_2\delta_1| \approx |\delta_2+\delta_1| \leq \epsilon$. So the maximum possible relative error is ϵ.
The only possible values if $c = 0$ are -Inf, Inf, and NaN, the NaN only occurs if $a \otimes b = 0$.

Exercise : 5.10

We can represent this math as $(a'-b')*2^E = (1.a_1a_2\cdots a_{52} - 1.b_1b_2\cdots b_{52})*2^E = c'*2^E$ or $a'-b' = c'$ where $a', b' \in [1, 1.\{1\}^{52}]$ and so $c' \in [1-1.\{1\}^{52}, 1.\{1\}^{52}-1] = [-0.\{1\}^{52}, 0.\{1\}^{52}]$. The exact value of $a'-b'$ clearly can't have more digits after the decimal than a' or b' so we can write $a'-b' = \pm 0.d_1d_2\cdots d_{52}$. Thus $a-b = \pm 0.d_1d_2\cdots d_{52} * 2^E$ examining this number we can clearly see that it can be expressed as a double without rounding.

Exercise : 5.15

(a) Using octave to add up the fractions I get $x = \frac{209715}{2097152}$

(b) This would simply be $|\frac{209715}{2097152} - 1/10|/(1/10) = 10*|\frac{209715-209715.2}{2097152}| = 10*\frac{.2}{2097152} = \frac{1}{1048576}$

(c) This would just be the absolute error times the time. $\frac{360000}{1048576} \approx 0.34332$sec.

(d) This would be $0.34332/3600 * 3750 \approx 0.35763$miles.