1 Determine the single precision 32-bit representation of the following decimal numbers:

a) 2^{-30}

$[2^{-30}]_{10}$
$= 1 \times 2^{-30}$
bias
1×2^{97}
$= [1.0 \times 10^{01100001}]$

0	01100001	00000000000000000000000

✓

b) 64.015625

$64_{10} = 0100\ 0000_2$

0.015625	
0.03125	0
0.0625	0
0.125	0
0.25	0
0.5	0
1.0	1

$= 0.000001_2$

$[64.015625]_{10}$
$= [01000000.000001]_2$
$= 1.000000000001 \times 10^{0110}$
bias
$1.000000000001 \times 10^{10000101}$

0	10000101	00000000000010000000000

✓

c) -8×2^{-24}

$[-8 \times 2^{-24}]_{10}$
bias
-8×2^{103}
$= [-1000.0 \times 10^{01101010}]_2$

$= -1.0 \times 10^{01100110}$

1	01101010	00000000000000000000000

✓

d) 0.5

$[0.5]_{10}$
$= [0.1]_2$
$= 1.0 \times 10^{-1}$
bias
1.0×10^{01111110}

0	01111110	00000000000000000000000

✓

e) 42.424242

5 sig. bits after 1

$42_{10} = 00101010$

0.424242	
0.848484	0
1.696968	1
1.393936	1
0.787872	0
1.575744	1
1.151488	1
0.302976	0
0.605952	0
1.211904	1
0.423808	0
0.847616	0
1.695232	1
1.390464	1
0.780928	0
1.561856	1
1.123712	1
0.247424	0
0.494848	0

need 18 (23-5) more bits for mantissa

$\rightarrow .01101100100111011 00$

$[42.424242]_{10}$

$= [00101010.01101100100111011 00]_2$

$= 1.01010011011001001101100 \times 10^{0101}$

bias

$1.01010011011001001101100 \times 10^{10000100}$

0	10000100	01010011011001001101100

✓

f) $76.234567 \times 10^{-15} \rightarrow [0.000000000000076234567]$

Goal: Rewrite in scientific notation w/ base 2

$76.234567 \times 10^{-15}$

$= 7.6234567 \times 10^{-14}$

Find Exponent

$2^K \leq 7.6234567 \times 10^{-14} < 2^{K+1}$

$\Rightarrow K = \lfloor \log_2(7.6234567 \times 10^{-14}) \rfloor$

$= \lfloor -43.57\ldots \rfloor$

$= -44$

$\therefore 2^{-44} \leq 7.6234567 \times 10^{-14} < 2^{-43}$

$5.68\ldots e\text{-}14 \leq 7.62\ldots e\text{-}14 < 1.13\ldots e\text{-}13$

$|2^{-44} - 7.62\ldots e\text{-}14| < |2^{-43} - 7.62\ldots e\text{-}14|$

$\therefore 7.6234567 \times 10^{-14} = C \times 2^{-44}$

Find Coefficient

$C = 2^{44} \times (7.6234567 \times 10^{-14})$

$= 1.341132685679949652787 2$

Coefficient to Binary

0.34...	
0.68...	0
1.36...	1
0.72...	0
1.45...	1
0.91...	0
1.83...	1
1.66...	1
1.32...	1
0.65...	0
1.31...	1
0.63...	0
1.27...	1
0.55...	0
1.11...	1
0.23...	0
0.47...	0
0.94...	0
1.88...	1
1.77...	1
1.54...	1
1.09...	1
0.18...	0
0.37...	0

bias

$[1.341132685679949652787 2 \times 2^{83}]_{10}$

$= 1.01010111010101000111100 \times 10^{01010011}$

0	01010011	01010111010101000111100

✓

g) 1.4345678

0 sig bits

$1.0_{10} = 1.0_2$

need 23 mo'

0.4345678	
0.8691356	0
1.7382712	1
1.4765424	1
0.9530848	0
1.9061696	1
1.8123392	1
1.6246784	1
1.2493568	1
0.4987136	0
0.9974272	0
1.9948544	1
1.9897088	1
1.9794176	1
1.9588352	1
1.9176704	1
1.8353408	1
1.6706816	1
1.3413632	1
0.6827264	0
1.3654528	1
0.7309056	0
1.4618112	1
0.9236224	0

$= 0.01101111001111111101010$

$1.01101111001111111101010 \times 10^{00000000}$

bias

$1.01101111001111111101010 \times 10^{01111111}$

0	01111111	01101111001111111101010

✓

1 bit off. why?

h) 3.14159265358979323846264338327950288 4

(my calculator only has 15 sig digits)

$[3]_{10}$

$= [0011]_2$

0.14...	
0.28...	0
0.56...	0
1.13...	1
0.26...	0
0.53...	0
1.06...	1
0.12...	0
0.24...	0
0.49...	0
0.99...	0
1.98...	1
1.96...	1
1.92...	1
1.85...	1
1.70...	1
1.41...	1
0.83...	0
1.66...	1
1.32...	1
0.65...	0
1.31...	1
0.63...	0

$[3.14...]_{10}$

$= [11.00100100001111110110101]_2$

$= 1.10010010000111111011010 \times 10$

0	10000000	10010010000111111011010

✓

1 bit off. why?

i) 3.14

$[3]_{10}$
$= [11]_2$

.14	
.28	0
.56	0
✗ 1.12	1
.24	0
.48	0
.96	0
✗ 1.92	1
✗ 1.84	1
✗ 1.68	1
✗ 1.36	1
.72	0
✗ 1.44	1
.88	0
✗ 1.76	1
✗ 1.52	1
✗ 1.04	1
.08	0
.16	0
.32	0
.64	0
✗ 1.28	1
0.56	0
✗ 1.12	1

$[3.14]_{10}$
$= [11.001000111101011100010]_2$
$= 1.1001000111101011100010 \times 10$

0	10000000	10010001111010111000010

✓

↑ 1 bit off. why?

j) 17/31

17/31	
34/31	1
6/31	0
12/31	0
24/31	0
48/31	1
34/31	1

repeats (5 times fit)

$[17/31]_{10}$
$= [0.10000110000110000110000110000]_2$
$= 1.00001100001100001100001100001 \times 10^{-01}$
bias
$= 1.00001100001100001100001100001 \times 10^{01111110}$

0	01111110	00001100001100001100001100

✓

↑ 1 bit off. why?

2 Identify the floating point numbers corresponding to the following single precision machine binaries:

a) 0 00000000 00000000000000000000000

0 ✓

b) 1 11111111 00000000000000000000000

-Infinity ✓

c) 0 10000001 01100000000000000000000

+ 129_{10} 1.011_2

-127

$= [1.011 \times 10^{10}]_2$

$= [2.375 \times 2^2]_{10}$

$= 9.25$ ✓

d) 0 01111111 00000000000000000000000

+ 127 1.0

-127

$= [1.0 \times 10^{00}]_2$

$= [2 \times 2^0]_{10}$

$= 2$ ✓

e) 0 01111011 00110011001100110011100

+ 123 1.1001...100

-127

8388608

4194304

$[1.10011001100110011001100 \times 10^{-0100}]_2$

$= 0.00011001100110011001100$

$= \left[\frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{512} + \frac{1}{4096} + \frac{1}{8192} + \frac{1}{65536} + \frac{1}{131072} + \frac{1}{1048576} + \frac{1}{2097152} + \frac{1}{16777216} + \frac{1}{33554432}\right]_{10}$

$= \frac{3}{32} + \frac{3}{512} + \frac{3}{8192} + \frac{3}{131072} + \frac{3}{2097152} + \frac{3}{33554432}$

$= 3\left(\frac{2^{20}+2^{16}+2^{12}+2^8+2^4+1}{2^{25}}\right)$ ✓

3 Consider a decimal machine in which floating-point numbers are represented with a precision of 12 decimal places. Compute the relative error for the following numbers (assume numbers are rounded correctly):

a) 1.56734565455567890621

x

$\hat{x} = 1.56734565456$

$\varepsilon = \left|\frac{x - \hat{x}}{x}\right|$

$= 2.0480485531 \times 10^{-13}$

b) $1 - 2^{-9}$

x

$\hat{x} = 1 - 2^{-9}$

$\varepsilon = \left|\frac{x - \hat{x}}{x}\right|$

$= 0$

c) $1 - 10^{-13}$

$x = 0.9999999999999$

$\hat{x} = 1$

$\varepsilon = \left|\frac{x - \hat{x}}{x}\right|$

$= 1.0000000000001$

d) $2.3456 - 0.000004567890111234$

$x = 2.345595432109887766$

$\hat{x} = 2.34559543211$

$\varepsilon = \left|\frac{x - \hat{x}}{x}\right|$

$\approx 4.7900234011 \times 10^{-14}$

e) $3.145623456789 12 - 3.145623451233476$

$x = 5.555644 \times 10^{-9}$

$\hat{x} = 5.556 \times 10^{-9}$

$\varepsilon = \left|\frac{x - \hat{x}}{x}\right|$

$\approx 0.00006407897998626$