Question 1

So, sign is - because of the first bit. Then, we flip the bits, add 1 and we add up the power of twos to get the number in decimal representing the magnitude. Combine with the negative symbol, and we have our number!

1. Flip the bits

$$11111111111111111111111111101010 = 00000000000000000000000000010101$$

2. Add 1

$$00000000000000000000000000010101 + 1 = 00000000000000000000000000010110$$

3. Calculate the magnitude

$$2^4 + 2^2 + 2^1 = 22$$

4. Combine

$$-22$$

Thus, our answer is -22.

Question 2

Suppose we have a binary8 format which follows IEEE standard. In this format the width of the exponent field is 4, and the width of the fraction field is 3.

a.

The exponent bias should be 7.

b.

The machine espilon is

$$00111001 - 00111000 = \epsilon = 2^{-3} = 0.125$$

c.

Largest: $01110111 = 2^8 * (1 + 2^-1 + 2^-2 + 2^4) = 256 * 1.875 = 480$
Smallest: $00001000 = 1 * 2^{-7} = 0.0078125$

d.

No, not all integers. For example, 479 cannot be stored. Proof: We can store 480, as seen above. However, the smallest step under would be 0 1110 110 which is $2^8 * (1 + 2^-1 + 2^-2)_2 = 256 * 1.75 = 448$
There are no steps in between. Thus, since 479 is an integer in between -480 and 480 and it cannot be represented, we conclude that there exist an integer which cannot exist in between the smallest and largest number in this system.

e. INCORRECT - REVIEW

largest: $00000111 = 2^{-7} * (2^{-2} + 2^{-4} + 2^{-8}) = 0.002471923828125$
smallest: $00000001 = 2^{-7} * 0.0078125 = 0.0000610352$

f. INCORRECT - REVIEW

What are the two floating point numbers (neither is equal to 5) closest to 5?
5 is 0 1000 so

- largest step before: 0 1000 000 = 4
- smallest step after: 0 1000 100 = 6

g.

Given number -(10.1011)2 Round it to a binary8 number using the four rounding modes.

h.

Find a real x in the range of subnormal numbers such that

$$\frac{|round(x) - x|}{|x|} > 12\epsilon$$

Question 3

Are the following statements true or false? If a statement is true, give a proof and if it's false, give a counterexample. We assume no overflow occurs in the calculations and the rounding mode used can be any of the four rounding modes.

a.

If x is a finite floating point number, then $x \oplus x = 2x$

b.

If x and y are two finite floating point number, then $x \ominus y = -(y \ominus x)$.

Question 4

What are the values of the expressions

a.

$-\inf/0$

b.

$\inf/(-\inf)$

c.

$3 * \mathrm{NaN} - \mathrm{NaN}$

d.

$0/(-\mathrm{NaN})$