Floating Point Reference Sheet

*(using an 8-bit representation like pg.86 of CS324 book) – look at the examples there as well

*__Bolded words__ are key differences between Denormalized and Normalized

Denormalized

Sign	***Exponent (exp)***				***Frac***		
1/0	0	0	0	0	1/0	1/0	1/0

e = 0 (since exp always has 0's)

Bias = (2^(# of exp bits – 1)) – 1 = 2^(4-1) – 1= 7

E = 1- bias = 1 – 7 = -6

f = (Value of Frac As Int)/(2^(# of frac Bits)) = (Value of Frac as Int)/(2^3)

M = f

(sign) = -1 if Sign=1

(sign) = +1 if Sign=0

V = M x (2^E) x (sign) = M x (2^(-6)) x (sign)

Normalized

Sign	***Exponent (exp)***				***Frac***		
1/0	1/0	1/0	1/0	1/0	1/0	1/0	1/0

Exp field needs at least one '1' in it and at least one '0' in it (i.e. Range 0001 to 1110 in binary).

e = int value of exp field

Bias = (2^(# of exp bits – 1)) – 1 = 2^(3) – 1= 7

E = e - bias

f = (Value of Frac As Int)/(2^(# of frac Bits)) = (Value of Frac as Int)/(2^3)

M = f + 1 = ((Value of frac as Int)/(2^3)) + (2^3)/(2^3)

(sign) = -1 if Sign=1

(sign) = +1 if Sign=0

V = M x (2^E) x (sign)

Infinity (+oo / - oo)

Sign	***Exponent (exp)***				***Frac***		
1/0	1	1	1	1	0	0	0

NaN (Not a Number)

Sign	***Exponent (exp)***				***Frac***		
1/0	1	1	1	1	<------	1 anywhere	------->

Frac contains at least one '1' in it

pg. 90 Default is to Round to Even i.e. On tie round to nearest even number

Mode	***$1.40***	***$1.60***	***$1.50***	***$2.50***	***$-1.50***
Round-to-even	$1	$2	$2	$2	$-2

Example Denormalized
0 0000 010
e = 0
E = -6
f = (0b010 / (2^3)) = 2 / (2^3) = (2/8)
M = 2/8
Value = (2/8) x (2^(-6)) = (2/(2^3)) x (1/(2^6)) = (2/(2^9)) = (2/512) = (1/256)

Example Normalized
1 0111 001
e = 0b0111 = 7
E = e-7 = 7-7 = 0
f = (0b001/ (2^3)) = (1/8)
M = (1/8) + 1 = (1/8) + (8/8) = (9/8)
Value = (9/8) x (2^0) x (-1) = (-9/8)

Side Notes (pg.83-84)
Single-precision floating-point format (a float in C)
s = 1 bit
exp (k) = 8 bits
frac (n) = 23 bits
32 bits total

Double-precision floating-point format (a double in C)
s = 1 bit
exp (k) = 11 bits
frac (n) = 52 bits
64 bits total