CSCI 2330 – Floating Point Exercises

1. Using an 8-bit IEEE floating point representation (with k=4 exponent bits and 3 fractional bits), convert **00110100** into a decimal value.

2. Using the same 8-bit representation, convert **10000101** into a decimal value (working with a fraction here is advisable).

3. If **d** is a double, does (d < 0.0) imply ((d * 2) < 0.0)? (remember this is not true for ints)

4. Excluding infinity, what is the decimal value of the largest 32-bit IEEE floating point number? You should be able to write down an exact expression (unsimplified).

5. IEEE 754 encodes the exponent value **E** using an unsigned **exp** field from which a **bias** value is subtracted. An alternate approach would be to just make **exp** encode a signed number and get rid with the bias term. Is there a reason to prefer the **unsigned - bias** approach?

Hint: think about the natural ordering of binary values (e.g., 000, 001, 010, 011, ...) and what the ordering of corresponding IEEE values would be.