## Exercises 1

- The largest number that can be represented by a floating point system (2, 5, -10, 10) is 0.11111 x 2^10 which is equal to 992 (base 10). The largest value of n where n! <= 992 is 6 (since 6! = 720). 6! = 0.101101 x 2^10, which needs 6 digits. 5! = 0.1111 x 2^7. Hence the largest number we can represent is n=5.
  - **b)** 7 has exponent 3 in the base-2 system. Hence we have

$$2^{-12} = 2^{3-+} \rightarrow + = 15$$

70 has exponent 7. Hence its distance from the next largest floating point number is

8 (base 10) =  $0.1 \times 2^4$ . Since 8 is a number in this floating point system and we have x < 8 < y, to have the smallest y-x we must have that x, 8, and y are consecutive numbers. The distance from x to 8 is  $2^(3-t)$ . The distance from y to 8 is  $2^(4-t)$ . Hence we have that the smallest y-x is

2. a) -- 0 0.1 0.101 0.11 0.01 1.1 1.1 10 10.1 11 11.1

Negative numbers are a mirror image of the positive ones (didn't want to draw them lol) \*\*Not to scale

There are 25 elements in S.

OFL = largest positive number = 11.1

UFL = smallest positive number = 0.1

$$|f(x)-x| \leq 0.1$$

$$|f(x)-x| \geq 0.1$$

$$\frac{|f(x)-x|}{|x|} \leq 1 \implies \text{machine epsilon}$$

3. 1.5 x 10^8 = 0.100011110000110100011 x 2^28
This uses more than 20 bits. We round to
0.1000111100001101001 x 2^28
The length on the last bit of this mantissa represents 2^10.

75 (base 10) = 0.1001011 x 2^7

The length on the last bit of this mantissa represents 2.