1. Precision of a floating point number is defined by the exponent part in the floating point representation as per IEEE 754 standards. We can see the same by looking at the single precision and double precision numbers in which the error percentage of any number (suppose e^x), is greater in single precision (32-bit) representation as compared to double precision (64-bit) representation as in single precision floating point numbers no. of exponent bits are 8 while in double precision floating point number it is 11 bits thus providing much greater range of accuracy.

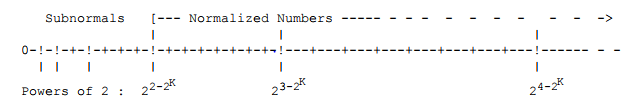
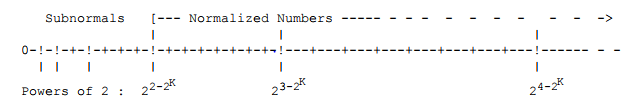
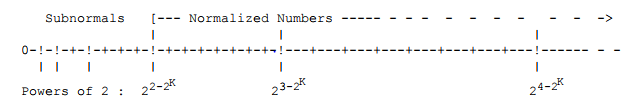
2. Normal and subnormal values as per IEEE standards:

For normal values, exponent part values goes from 0x01 to 0xFE while for subnormal values exponent part has all its bits as 0.

Normal values

Subnormal values

1

0

Basically, subnormal numbers gives the vast range for the numbers before zero which comes in handy for the converging series to zero as below 1 the number doesn’t abruptly falls to zero rather goes to much smaller number as intended before terminating to 0. It is also helpful in the underflow condition.

3. There are 5 methods for rounding off floating point numbers as per IEEE standards:

Round towards zero: In this the numbers are rounded towards zero for example 2.30 is rounded to 2, 5.53 is rounded to 5 and -4.6 is rounded to -4

Round down: In this the number is rounded towards negative infinity. For example 2.30 is rounded to 2, -4.6 is rounded to -5.

Round up: In this the number is rounded towards positive infinity. For example 2.30 is rounded to 3, -4.6 is rounded to -4.

Round-to-nearest: In this the number is rounded to the nearest integer value. But this case causes the problem with the numbers such as 1.5,3.5 as in these cases we don’t know where to round about it to.

Round-to-even: This is the default rounding mode for IEEE floating point numbers. In this the numbers are rounded to the nearest even numbers. For example 1.5 is rounded to 2, 2.5 is rounded to 2, -1.5 is rounded to -2.