Exercise 1. Sam measured the box to the nearest 2 cm, and got 24cm×24cm×20cm. Measuring to the nearest 2 cm means the true value could be up to 1 cm smaller or larger. Find upper bounds for the absolute and relative errors in the box volume.
Hint. Consider using first principles.

Exercise 2. Any human-made computer is a finite-state machine, which can thus store only a finite set of numbers. Nowadays, almost entirely all computers use floating-point arithmetic, FP, which supports a trade-off between range and precision, when approximating real numbers as rational numbers defined by

$$F\left(b,p,L,U\right)\equiv\left\{x\in\mathbb{Q}\,\middle|\,x=0\text{ or }x=\pm\left(\frac{d_1}{b^1}+\frac{d_2}{b^2}+\ldots+\frac{d_p}{b^p}\right)\times b^e\right\}$$

This set of FP numbers uses p significant digits, d_i, base $b\geq 2$, such that $0\leq d_i\leq b-1$, and range (L,U) with $L\leq e\leq U$. For uniqueness, it is typically assumed that $d_1\neq 0$. In such an event, d_1 is called the principle significant digit, while d_p is the last significant digit and the representation of x is called normalized.

Consider a simple normalized FP system, F(10,4,-1,2). If $x\in F$, let x^+ be the next representable number. Find the following.

1) The total number of elements in F i.e. the cardinality of this FP system.
2) The greatest number in F, x_{max}.
3) The smallest positive number in F, x_{min}.
4) The distance, Δ, between unity and the next representable number, $\Delta=1^+-1$
5) If chopping (truncation) method is used, what is the greatest absolute and relative error when representing numbers $1<x<1^+$ in this FP system?
6) If rounding to nearest is used, what is the greatest absolute and relative error when representing numbers $1<x<1^+$ in this FP system?

Exercise 3. The IEEE standard does not define the terms machine epsilon nor unit roundoff, so differing definitions of these terms are in use, which can cause some confusion. Most numerical analysts use the words *machine epsilon* and *unit roundoff* interchangeably following the definition given by Prof. James Demmel. Machine epsilon is a maximum relative error of the chosen rounding procedure. Machine epsilon is often written as ε or ε_M whereas unit roundoff is typically denoted by **u**.

Among others, the IEEE standard defines the 32-bit base-2 format also known as Binary32 or single precision requiring 23 bits to store the significand. The first significant digit in the binary normalized form is always one, $d_1 = 1$, thus there is no need to keep this digit (hidden bit) in the computer memory hence making effectively, $p = 24$.

Do the following.

1) Write a computer code to find $\Delta=1^+-1$ in Binary32 arithmetic, which can be defined as the smallest positive number Δ in Binary32 such that $1+\Delta>1$. Express Δ in terms of the precision p.
2) In analogy to Exercise 2, deduce the machine epsilon in Binary32 when truncation is used.
3) Find the machine epsilon when rounding to the nearest is used.

4) In both MATLAB and Python, all calculations involving floating point numbers are performed in double precision by default. Check if the condition $1+2^{-52}>1$ is true or false in Binary64 arithmetic.
5) Check if the condition $1+2^{-53}=1$ is true or false in Binary64 arithmetic.
6) Based on the results in 4) and 5), deduce the parameter p and the machine epsilons (note plural) for truncating/chopping and rounding to nearest methods.

Exercise 4. In the subtraction 37.593621 − 37.584216, how many bits of significance will be lost?

Exercise 5. Since any number used in calculations with a computer system must conform to the format of numbers in that system, it must have a finite expansion. Numbers that have a nonterminating expansion cannot be accommodated precisely. Moreover, a number that has a terminating expansion in one base may have a nonterminating expansion in another. A good example of this is the following simple fraction

$$\frac{1}{10}=\left(0.1\right)_{10}=\left(0.0\ 0011\ 0011\ 0011\ 0011\ 0011\ 0011\ \ldots\right)_2=0.0\overline{0011}_2$$

Evaluate the relative representation error of this fraction in Binary32 arithmetic. To this end, define, $h=\mathrm{fl}\left(0.1\right)$ in single precision. The representation error is, by definition,

$$\delta\equiv\frac{h-0.1}{0.1}=10\cdot h-1$$

1) Evaluate this quantity, $\delta=10\cdot h-1$, in double precision.
2) Verify that $\left|\delta\right|\le\varepsilon_M$

Exercise 6. The Scottish mathematician, James Stirling, 1692 – 1770, proved a formula

$$\lim_{n\to\infty}\frac{n!}{\sqrt{2\pi n}\left(n/e\right)^n}=1$$

that gives us an approximation for large factorials, Stirling's formula or approximation,

$$n!\approx\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Bounds for the above ratio can be given as

$$\forall n\in\mathbb{N}\quad 1\le\frac{n!}{\sqrt{2\pi n}\left(n/e\right)^n}\le\frac{e}{\sqrt{2\pi}}=1.08444...$$

Evaluate the actual absolute and relative errors for the approximation

$$9!\approx\sqrt{18\pi}\left(\frac{9}{e}\right)^9$$

Does it satisfy the above bounds?

Exercise 7. The Bessel functions $J_n(x)$ are defined by

$$J_n(x) \equiv \frac{1}{\pi}\int_0^{\pi} \cos(x\sin\theta - n\theta)\,d\theta$$

a) Establish that $|J_n(x)| \le 1$

b) It is known that

$$J_{n+1}(x) = \frac{2n}{x}J_n(x) - J_{n-1}(x)$$

Use this equation to compute $J_1(1)$, $J_2(1)$, …, $J_{20}(1)$, starting from known values $J_0(1) \approx 0.7651976865$ and $J_1(1) \approx 0.4400505857$. Keep in mind a).

c) Another recursive relation is

$$J_{n-1}(x) = \frac{2n}{x}J_n(x) - J_{n+1}(x)$$

Starting with the known values $J_{20}(1) \approx 3.873503009 \times 10^{-25}$ and $J_{19}(1) \approx 1.548478441 \times 10^{-23}$, use this equation to compute $J_{18}(1)$, $J_{17}(1)$, …, $J_0(1)$. Analyse the results. It is helpful to fill in the following table

n	***J_n*(1) forward**	***J_n*(1) backward**
0	**0.7651976865**	
1	**0.4400505857**	
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		**1.548478441e-23**
20		**3.873503009e-25**

Exercise 8. We are trying to find the smallest root of the equation

$$x^2 - 140x + 1 = 0$$

Which of the formulas

$$x_1 = 70 - \sqrt{4899} \qquad \text{or} \qquad x_1 = \frac{1}{70 + \sqrt{4899}}$$

gives more exact result when using 4-decimal-digit arithmetic? Investigate two rounding procedures, truncation to a 4-digit number and rounding to the nearest 4-digit number. Compare your results with the “exact” root, i.e. evaluated in double precision.

Exercise 9. Find the third Taylor polynomial for the function $f(x) = \sqrt{1+x}$ about $x_0 = 0$, i.e. find the Taylor expansion of the third order

$$f(x) = P_3(x) + E_3(x) = f(0) + f'(0)x + \frac{f''(x)}{2!}x^2 + \frac{f'''(x)}{3!}x^3 + E_3(x)$$

Clearly, $\sqrt{x} = f(x-1) \approx P_3(x-1)$. Approximate $\sqrt{0.5}$, $\sqrt{0.75}$, $\sqrt{1.25}$ and $\sqrt{1.5}$ using $P_3(x)$ and fill in the table.

x	$\sqrt{x}$	$P_3(x\text{-}1)$	errAbs	errRel
0.5				
0.75				
1.25				
1.5				

Exercise 10. The hyperbolic sine function is defined as

$$\sinh(x) \equiv \frac{e^x - e^{-x}}{2}$$

a) Prove that the inverse hyperbolic function is given by

$$\sinh^{-1}(x) = \ln\left(x + \sqrt{x^2+1}\right)$$

b) Show how to avoid loss of significance in computing $\sinh^{-1}(x)$ when x is a small negative number. Hint: find and exploit the relationship between $\sinh^{-1}(x)$ and $\sinh^{-1}(-x)$.