ME209 - Numerical Methods

# Lecture 2: Chapter 3 – Approximations and Round-Off Errors

By:

Prof. Dr. Ahmet Erkliğ

# Significant Figures

- Whenever we employ a number in a computation, we must have assurance that it can be used with confidence.

Visual inspection of the speedometer indicates that the car is traveling between <u>48 and 49 km/h</u>.

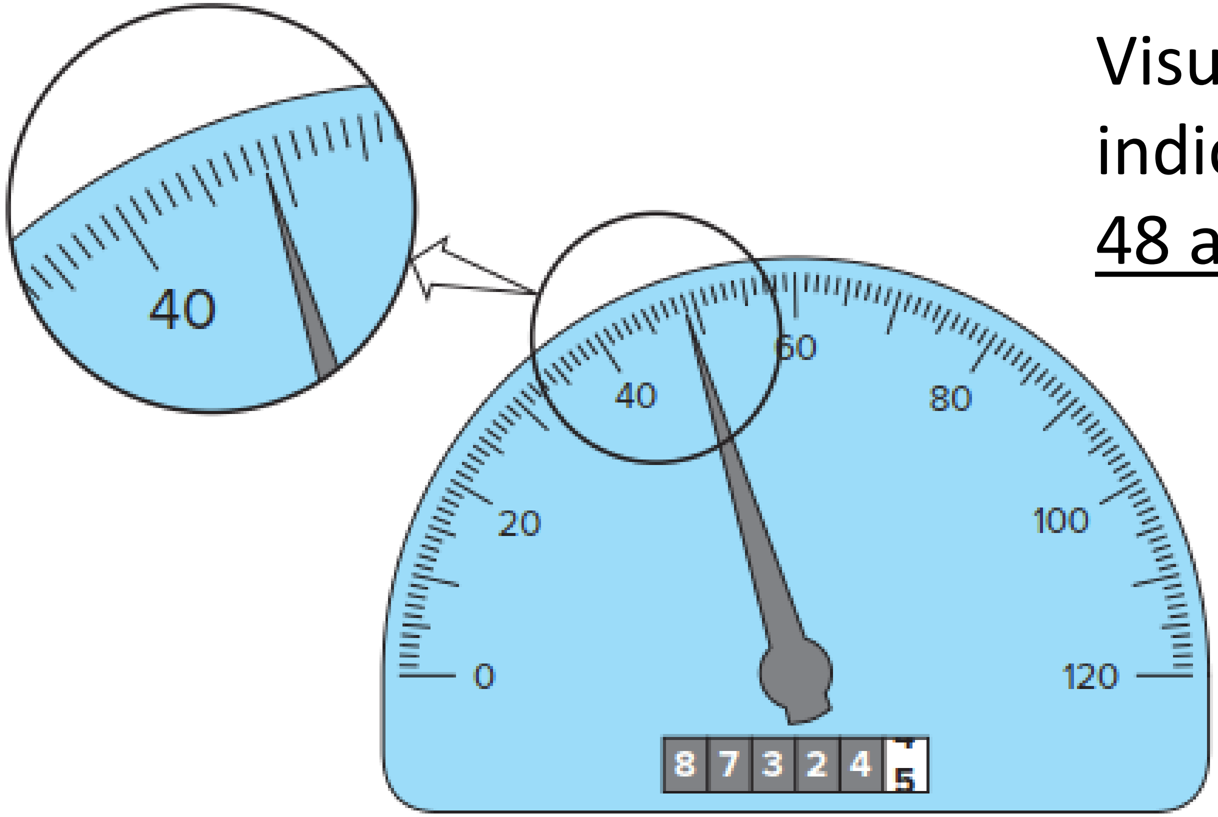
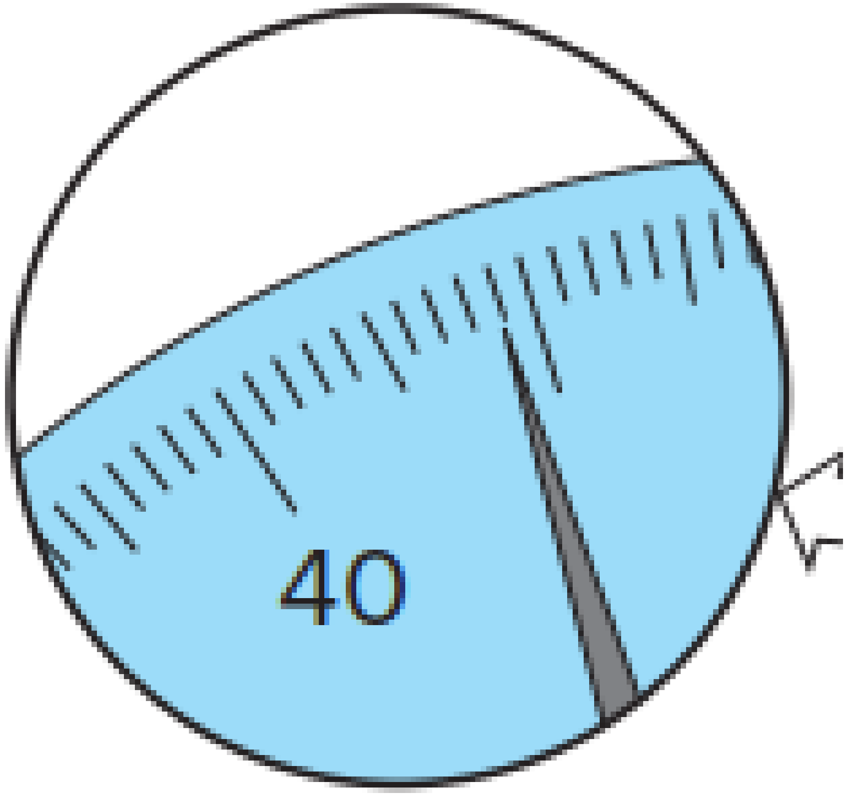



**FIGURE 3.1**
An automobile speedometer and odometer illustrating the concept of a significant figure.

# Significant Figures

We insist that the speed be estimated to one decimal place. For this case, one person might say 48.8, whereas another might say 48.9 km/h.




Therefore, because of the limits of this instrument, only the first two digits can be used with confidence. Estimates of the third digit (or higher) must be viewed as approximations.

**FIGURE 3.1**
An automobile speedometer and odometer illustrating the concept of a significant figure.

# Significant Figures

In contrast, the odometer provides up to six certain digits. From Fig. 3.1, we can conclude that the car has traveled slightly less than 87,324.5 km during its lifetime. In this case, the seventh digit (and higher) is uncertain.

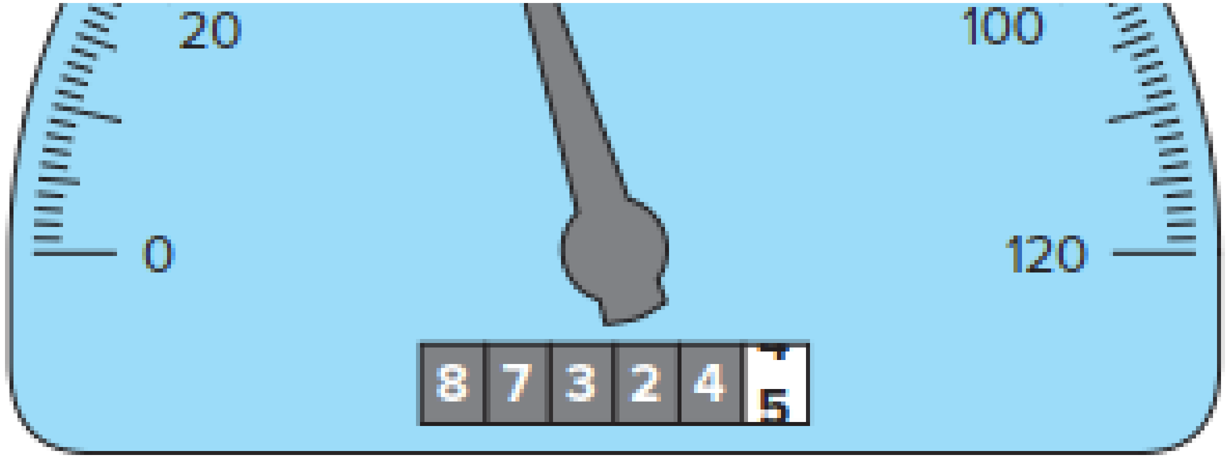



**FIGURE 3.1**
An automobile speedometer and odometer illustrating the concept of a significant figure.

# Significant Figures

- The concept of a significant figure, or digit, has been developed to formally designate the reliability of a numerical value.
- The ***significant digits*** of a number are those that can be used with confidence. They correspond to the **number of certain digits plus one estimated digit**.
- For example, the speedometer and the odometer in Fig. 3.1 yield readings of three and seven significant figures, respectively
- For the speedometer: three significant figures: 48.5
- For the odometer: seven-significant-figure reading of 87,324.45

# Significant Figures

- Although it is usually a straightforward procedure to ascertain the significant figures of a number, some cases can lead to confusion.
- For example, zeros are not always significant figures because they may be necessary just to locate a decimal point.
- The numbers 0.00001845, 0.0001845, and 0.001845 all have four significant figures.
- For example, at face value the number 45,300 may have three, four, or five significant digits, depending on whether the zeros are known with confidence.
- Such uncertainty can be resolved by using scientific notation, where $4.53 \times 10^4$, $4.530 \times 10^4$, $4.5300 \times 10^4$ designate that the number is known to three, four, and five significant figures, respectively.

# Significant Figures

The concept of significant figures has two important implications for our study of numerical methods:

- As introduced in the falling parachutist problem, numerical methods yield approximate results. We must, therefore, develop criteria to specify how confident we are in our approximate result.
- Although quantities such as π, e, or √7 represent specific quantities, they cannot be expressed exactly by a limited number of digits.

  $\pi = 3.141592653589793238462643\ \ldots$

# Example

If we are measuring the lengths of the base and height of a triangle with an instrument that provides values to the nearest tenth of an inch, then the measured lengths of b = 12.3 in and h = 17.2 in are expressed to three significant digits.

Calculate area with three significant figures.

**$A = 0.5bh = 0.5(12.3)(17.2) = 106\ in^2$**

# Accuracy and Precision

The errors associated with both calculations and measurements can be characterized with regard to their accuracy and precision.

- **Accuracy** refers to how closely a computed or measured value agrees with the true value.
- **Precision** refers to how closely individual computed or measured values agree with each other.
- **Inaccuracy** (also called bias) is defined as systematic deviation from the truth.
- **Imprecision** (also called uncertainty), on the other hand, refers to the magnitude of the scatter.

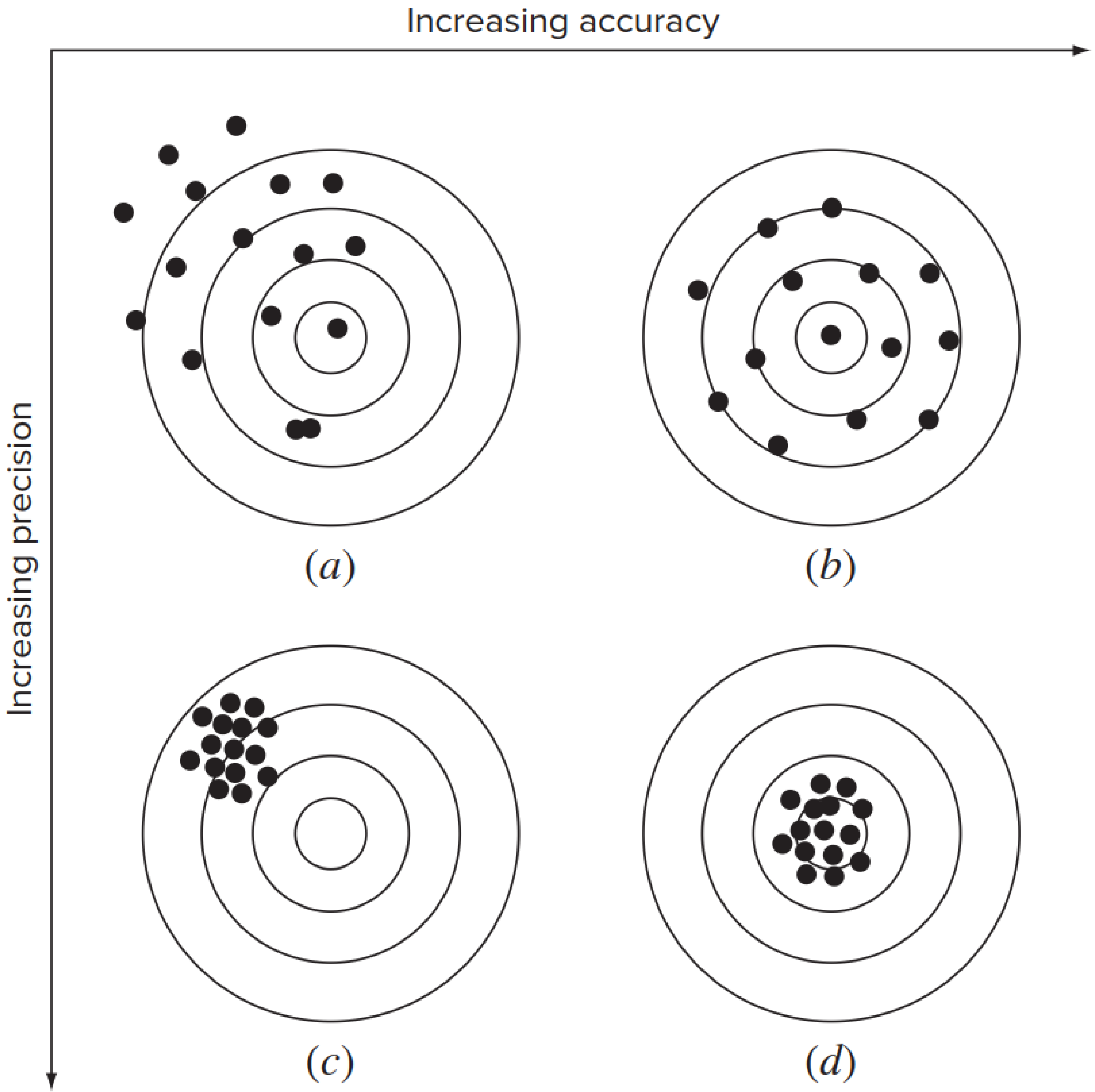
# Accuracy and Precision

(a) Inaccurate and imprecise;

(b) accurate and imprecise;

(c) inaccurate and precise;

(d) accurate and precise.



# Error Types

An error in estimating or determining a quantity of interest can be defined as a deviation from its unknown true value.

In general, errors can be classified based on their sources as non-numerical and numerical errors.

- **Non-numerical errors** include (1) modeling errors, (2) blunders and mistakes, and (3) uncertainty in information and data.
- **Numerical errors** include (1) round-off errors, (2) truncation errors, (3) propagation errors, and (4) mathematical-approximation errors.

# Error Definitions

Numerical errors arise from the use of approximations to represent exact mathematical operations and quantities.

- **Truncation errors,** when approximations are used to represent exact mathematical procedures,
- **Round-off errors,** when numbers having limited significant figures are used to represent exact numbers.

The relationship between the exact, or true, result and the approximation can be formulated as

$$\boxed{\text{True value} = \text{approximation} + \text{error}}$$

We can define the error as the difference between the computed and true values of a number:

$$\boxed{E_t = \text{True value - approximation}}$$

A relative true error ($\varepsilon_t$) is defined as the true error $E_t$ relative to the true value:

$$\boxed{\varepsilon_t = \frac{E_t}{True\,Value}}$$

The relative true error could be expressed as a percentage by multiplying $\varepsilon_t$ by 100·

$$\boxed{\varepsilon_t = \frac{E_t}{True\,Value} 100\%}$$

**Problem Statement.** Suppose that you have the task of measuring the lengths of a bridge and a rivet and come up with 9999 and 9 cm, respectively. If the true values are 10,000 and 10 cm, respectively, compute **(*a*)** the true error and **(*b*)** the true percent relative error for each case.

**Solution.**

**(a)** The error for measuring the bridge is $E_t = 10{,}000 - 9999 = 1$ cm

and for the rivet it is $E_t = 10 - 9 = 1$ cm

**(b)** The percent relative error for the bridge is

$$\varepsilon_t = \frac{1}{10{,}000}100\% = 0.01\%$$

for rivet is

$$\varepsilon_t = \frac{1}{10}100\% = 10\%$$

- For numerical methods, the true value will be known only when we deal with functions that can be solved analytically.
- However, in real-world applications, we will obviously not know the true answer *a priori*. For these situations, an alternative is to normalize the error using the best available estimate of the true value, that is, to **the approximation** itself,

Error can be written as $e_i = x_i - x_t$

where $e_i$ is the error in $x$ at iteration $i$, and $x_i$ is the computed value of $x$ from iteration $i$; similarly, the error for iteration $(i + 1)$ is

$$e_{i+1} = x_{i+1} - x_i$$

the change in the error

$$\Delta e_i = e_{i+1} - e_i = (x_{i+1} - x_t) - (x_i - x_t) = x_{i+1} - x_i$$

## Example

$$x^3 - 3x^2 - 6x + 8 = 0 \quad \Rightarrow \quad x = \sqrt{3x + 6 - \frac{8}{x}}$$

If we choose $x_0 = 2$ as the initial estimate of the positive value of x,

$$x_1 = \sqrt{3x_0 + 6 - \frac{8}{x_0}} = 2.828427 \quad \Rightarrow \quad e_1 = x_1 - x_0 = 0.828427$$

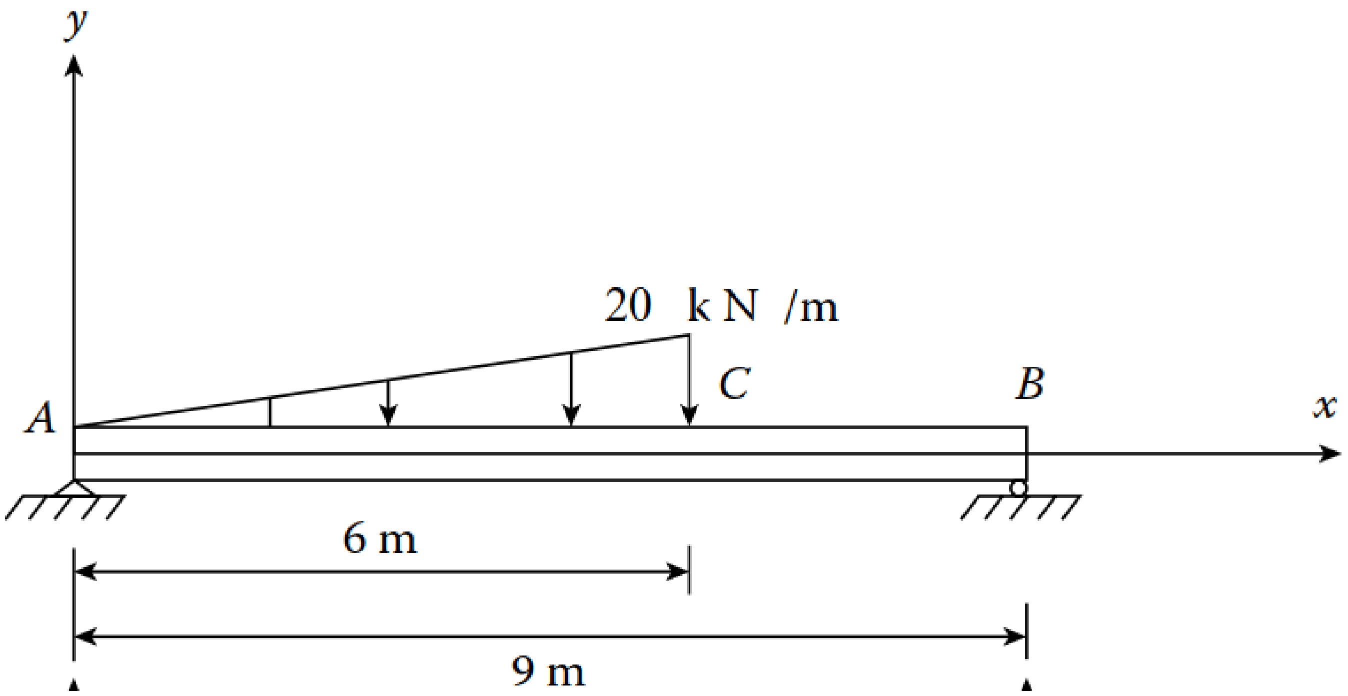
$x_2 = 3.414214$ $\qquad$ $e_2 = 0.585786$

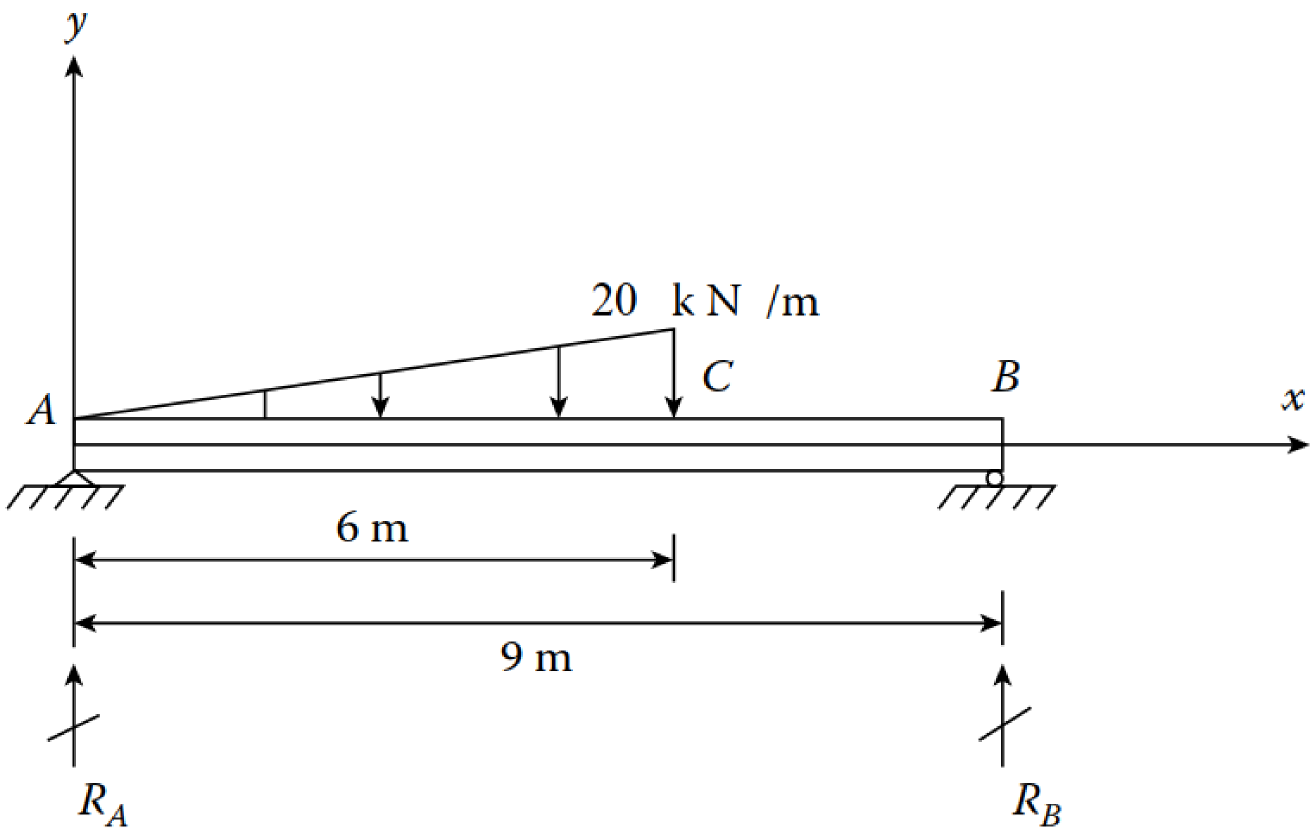
If the tolerable error were set at 0.025

| **Trial *i*** | $x_i$ | $\Delta e_i$ | $x_i - x_i$ |
|---|---|---|---|
| 0 | 2.000000 | – | 2.000000 |
| 1 | 2.828427 | 0.828427 | 1.171573 |
| 2 | 3.414214 | 0.585786 | 0.585786 |
| 3 | 3.728203 | 0.313989 | 0.271797 |
| 4 | 3.877989 | 0.149787 | 0.122011 |
| 5 | 3.946016 | 0.068027 | 0.053984 |
| 6 | 3.976265 | 0.030249 | 0.023735 |
| 7 | 3.989594 | $\boxed{0.013328}$ | 0.010406 |
| 8 | 3.995443 | 0.005849 | 0.004557 |
| 9 | 3.998005 | 0.002563 | 0.001995 |
| 10 | 3.999127 | 0.001122 | 0.000873 |

# BENDING MOMENT FOR A BEAM

The maximum-moment location corresponds to the point of zero shear force. Numerical methods can be used to determine this point along the span of the beam.






$\Sigma M_A = 0$

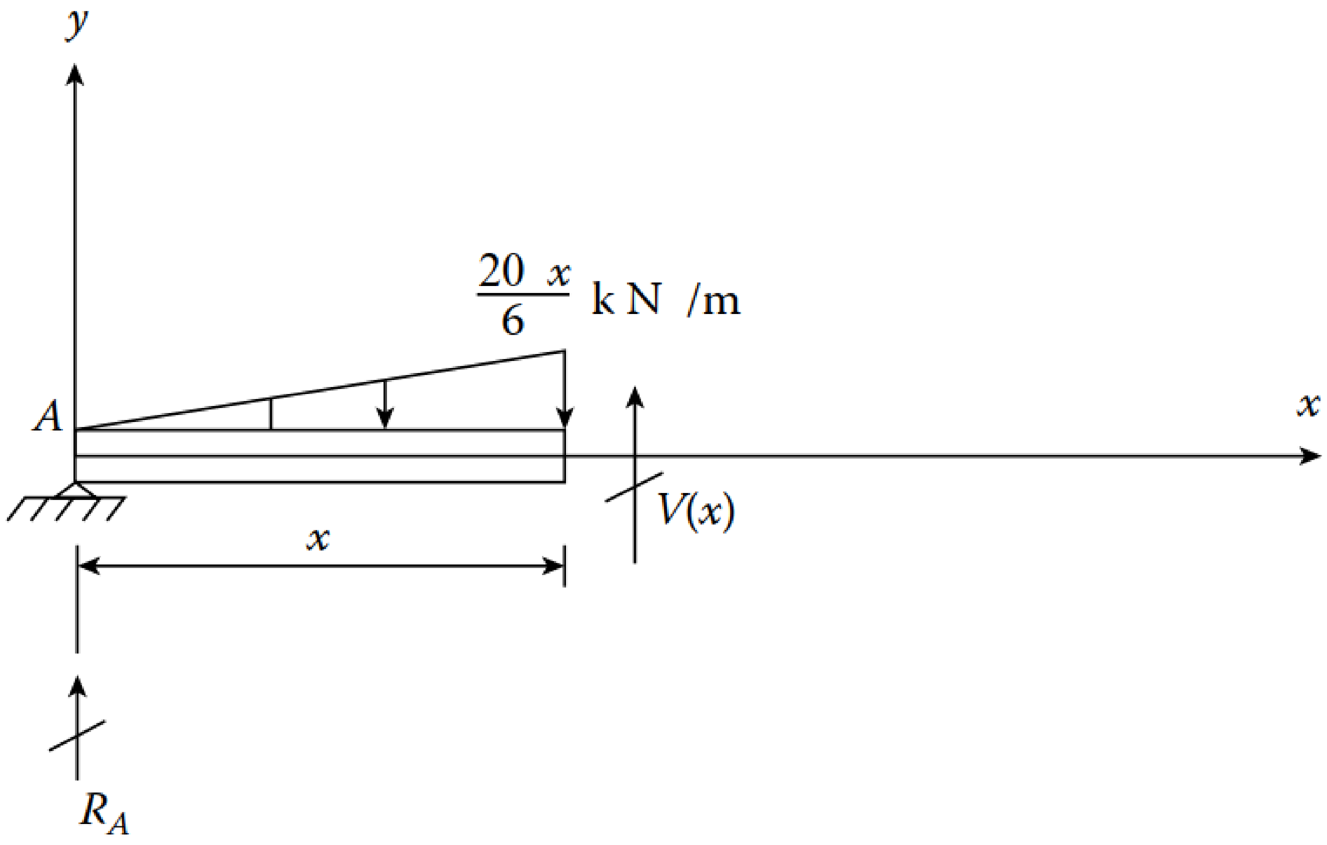
$\Sigma F_Y = 0$

$R_B = 26.6667^{+}$ kN

$R_A = 33.3333^{+}$ kN

The numerical solution starts by assuming a value for the location of the maximum bending moment—that is, zero shear force ($V$) location, say $x_0$.

$$V(x_0) = R_A - \frac{1}{2} x_0 \frac{x_0}{6}(20)$$



The shear force $V(x_0)$ based on this initial estimate of $x_0$ might not be zero· Then a new value for the location (distance), $x_{i+1}$, can be computed based on $x_i$ (with a starting value of $x_0$) by adjusting the original distance using the following adjustment ($\Delta x$):

$$\Delta x = \frac{R_A - \frac{1}{2} x_i \frac{x_i}{6}(20)}{R_A}$$

Therefore, the new distance is

$$x_{i+1} = x_i + \Delta x = x_i + \frac{R_A - \frac{1}{2} x_i \frac{x_i}{6}(20)}{R_A}$$

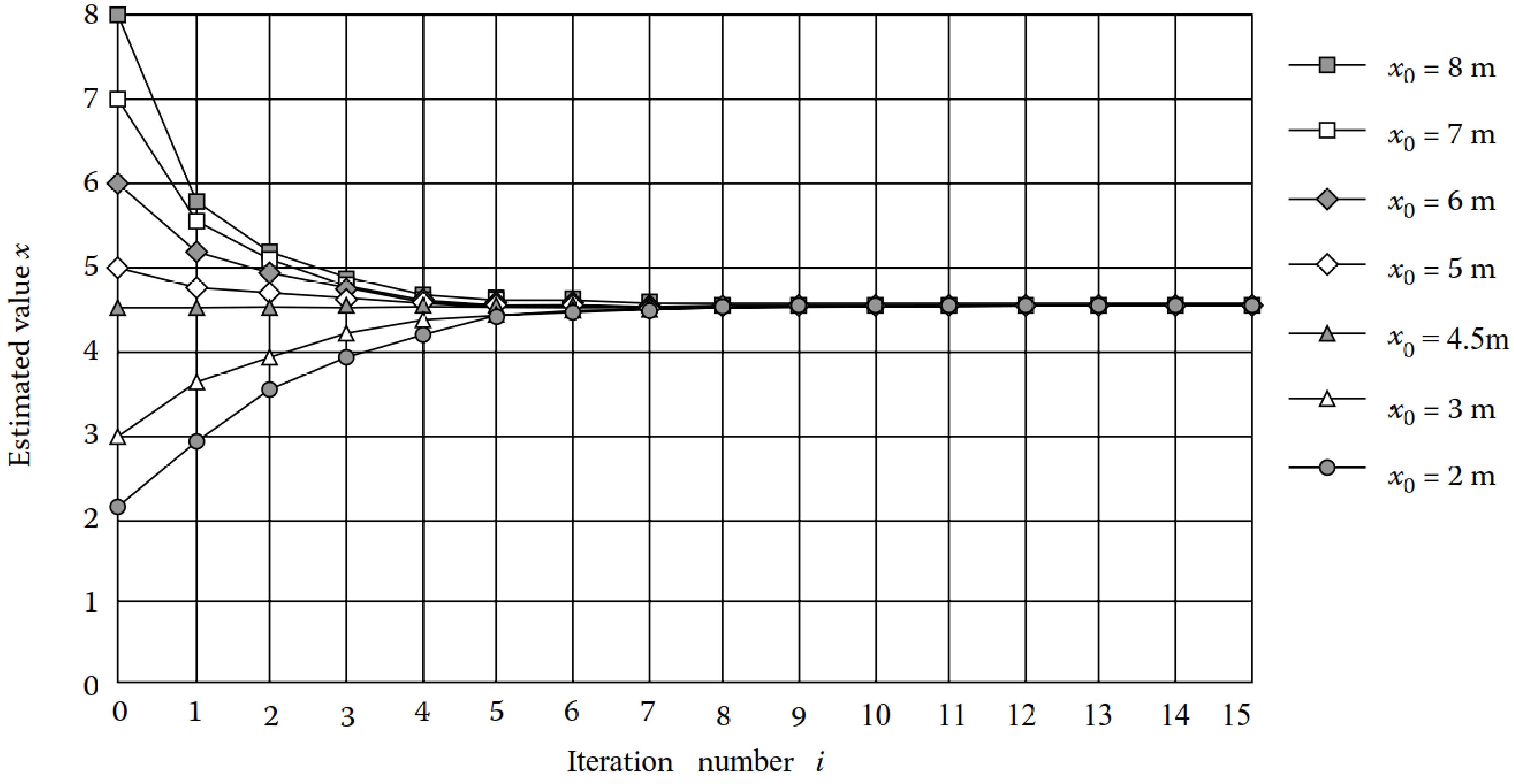
This process should be continued until some tolerance level (error level) is achieved.

The exact or true distance ($x_i$) can be determined by solving for the distance from $V(x_0) = 0$, which is

$$x_i = \sqrt{2R_A \frac{6}{20}} = \sqrt{20} = 4.4721359 \text{ m}$$

## Iterations with an Initial Value of 4.5 m

| Iteration *i* | $x_i$ | Error (%) |
|---|---|---|
| 0 | 4.5 | – |
| 1 | 4.4875 | 0.34355 |
| 2 | 4.480617 | 0.189646 |
| 3 | 4.476821 | 0.104753 |
| 4 | 4.474725 | 0.057882 |
| 5 | 4.473567 | 0.031989 |
| 6 | 4.472927 | 0.017681 |
| 7 | 4.472573 | 0.009773 |
| 8 | 4.472378 | 0.005402 |
| 9 | 4.472269 | 0.002986 |
| 10 | 4.47221 | 0.001651 |
| 11 | 4.472177 | 0.000912 |
| 12 | 4.472159 | 0.000504 |
| 13 | 4.472148 | 0.000279 |
| 14 | 4.472143 | 0.000154 |
| 15 | 4.47214 | 8.52E-05 |



Estimated value $x$
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Iteration number $i$
$x_0$ = 8 m
$x_0$ = 7 m
$x_0$ = 6 m
$x_0$ = 5 m
$x_0$ = 4.5m
$x_0$ = 3 m
$x_0$ = 2 m

# TAYLOR SERIES EXPANSION

A Taylor series is commonly used as a basis of approximation in numerical analysis.

A Taylor series is the sum of functions based on continually increasing derivatives. For a function $f(x)$ that depends on only one independent variable $x$, the value of the function at point $x_0 + h$ can be approximated by the following Taylor series:

$$f(x_0+h)=f(x_0)+hf^{(1)}(x_0)+\frac{h^2}{2!}f^{(2)}(x_0)+\frac{h^3}{3!}f^{(3)}(x_0)$$

$$+\cdots+\frac{(h^n)}{n!}f^n(x_0)+R_{n+1}$$

in which

- $x_0$ is some base value (or starting value) of the independent variable $x$
- $h$ is the distance between $x_0$ and the point $x$ at which the value of the function is needed—that is, $h = x - x_0$
- $n!$ is the factorial of $n$ (i.e., $n! = n(n - 1)(n - 2) \dots 1$)
- $R_{n+1}$ is the remainder of the Taylor series expansion
- The superscript $(n)$ on the function $f^{(n)}$ indicates the $n$th derivative of the function $f(x)$

$$f(x_0+h)=\sum_{k=0}^{\infty}\frac{h^k}{k!}f^{(k)}(x_0)$$

# TAYLOR SERIES EXPANSION OF THE SQUARE ROOT

The square-root function, using the Taylor series expansion, can be expressed as

$$f(x) = \sqrt{x}$$

To evaluate the Taylor series, the derivatives of the function are developed

$$f^{(1)}(x) = \frac{1}{2} x^{-0.5} \qquad f^{(2)}(x) = -\frac{1}{4} x^{-1.5} \qquad f^{(3)}(x) = \frac{3}{8} x^{-2.5}$$

For a base point $x_0$ = 1 and $h$ = 0.001, the four terms of the Taylor series produce the fol-lowing estimate for the square root of 1.001:

$$f(1.001) = \sqrt{1.001} \approx \sqrt{1} + 0.5(0.001)(1)^{-0.5} - \frac{1}{4(2!)}(0.001)^2(1)^{-1.5}$$
$$+ \frac{3}{8(3!)}(0.001)^3(1)^{-2.5}$$

$$f(1.001) \approx 1 + 0.5 \times 10^{-3} - 0.125 \times 10^{-6} + 0.625 \times 10^{-10} = 1.0004999$$

# EXAMPLE SERIES

The exponential evaluation to the base *e* of *x* can be expressed by the following series:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

The natural logarithm of *x* can be expressed using a Taylor series as

$$\ln(x) = (x-1) - \frac{(x-1)^2}{2!} + \frac{(x-1)^3}{3!} - \cdots = \sum_{k=0}^{\infty} (-1)^{k-1} \frac{(x-k)^k}{k!} \qquad \text{for } 0 < x \leq 2$$

The sine and cosine functions can also be expressed using the Taylor series as

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}$$

- For numerical methods, the true value will be known only when we deal with functions that can be solved analytically.
- However, in real-world applications, we will obviously not know the true answer *a priori*. For these situations, an alternative is to normalize the error using the best available estimate of the true value, that is, to **the approximation** itself,

$$\varepsilon_a = \frac{\text{approximate error}}{\text{approximation}} 100\%$$

where the subscript $a$ signifies that the error is normalized to an approximate value.

$$\varepsilon_a = \frac{\text{current approximation} - \text{previous approximation}}{\text{current approximation}} 100\%$$

**Tolerance:** Many numerical methods work in an iterative fashion. There should be a stopping criteria for these methods. We stop when the error level drops below a certain tolerance value ($\varepsilon_s$) that we select ($|\varepsilon_a| < \varepsilon_s$)

# Error Estimates for Iterative Methods

**Problem Statement.** In mathematics, functions can often be represented by infinite series. For example, the exponential function can be computed using **Maclaurin series expansion**

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

Starting with the simplest version, $e^x = 1$, add terms one at a time to estimate $e^{0.5}$. After each new term is added, compute the true and approximate percent relative errors. Note that the true value is $e^{0.5} = 1.648721\ldots$. Add terms until the absolute value of the approximate error estimate $\varepsilon_a$ falls below a prespecified error criterion $\varepsilon_s$ conforming to three significant figures.

**Solution.** First, determine the error criterion that ensures a result is correct to at least three significant figures:

$\varepsilon_s = (0.5 \text{ x } 10^{2-3})\% = 0.05\%$

Thus, we will add terms to the series until $\varepsilon_a$ falls below this level.

The first estimate is equal to 1. The second estimate is then generated by adding the second term, as in

$$e^x = 1 + x$$

or for $x = 0.5$,

$$e^{0.5} = 1 + 0.5 = 1.5$$

This represents a true percent relative error of

$$\varepsilon_t = \frac{1.648721 - 1.5}{1.648721} 100\% = 9.02\%$$

An approximate estimate of the error

$$\varepsilon_a = \frac{1.5 - 1}{1.5} 100\% = 33.3\%$$

Because $\varepsilon_a$ is not less than the required value of $\varepsilon_s$, we would continue the computation by adding another term, $x^2/2!$, and repeating the error calculations. The process is continued until $\varepsilon_a < \varepsilon_s$

| Terms | Result | $\varepsilon_t$ (%) | $\varepsilon_a$ (%) |
|---|---|---|---|
| 1 | 1 | 39.3 | |
| 2 | 1.5 | 9.02 | 33.3 |
| 3 | 1.625 | 1.44 | 7.69 |
| 4 | 1.645833333 | 0.175 | 1.27 |
| 5 | 1.648437500 | 0.0172 | 0.158 |
| 6 | 1.648697917 | 0.00142 | 0.0158 |