PROBLEM-SOLVING TUTORIAL

# 6.1 Data Analysis: Tutorial

## Introduction

In this chapter we cover some tools that are used in all areas of science: the tools of data analysis. Since you may not be familiar with the concepts, we will teach some of the basics of data analysis as we introduce those features in Mathcad. Once you see how easy Mathcad can summarize and display the results of an experiment, you will realize that analyzing data without a computer is just unnecessary punishment. In this Tutorial, we cover the statistics that are used to describe a set of data.

After having done an experiment, you must be able to describe your results. When reporting the value of a quantity, such as speed for example, you must also give an estimate of the error associated with the quantity. For example, two scientists do separate experiments to determine the maximum flying speed of the South African swallow:

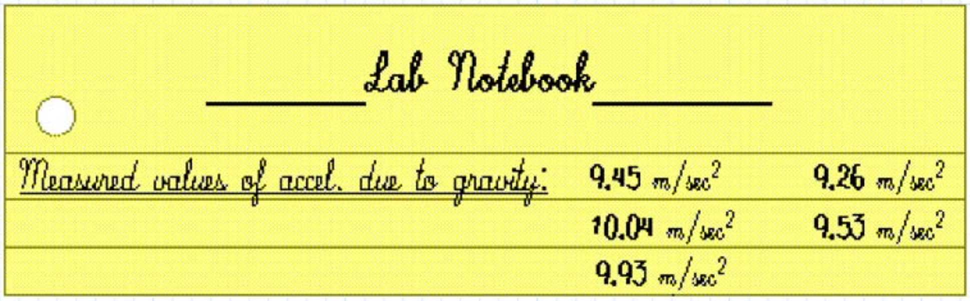
| Madame Curie | Einstein |
|---|---|
| 14.5 mph ± **0.2 mph** | 14.5 mph ± **10 mph** |

Although the reported values of the speeds are the same, Madame Curie is much more confident of her result than than Einstein since her reported error in the speed is much less than Einstein's. Notice that the speed reported by Einstein could be anywhere from $24.5 \cdot mph$ to $4.5 \cdot mph$ !

In this chapter, you will learn how Mathcad can be used to summarize your data and estimate the error in your experimental results.

## Experiment

Say we want to know the acceleration due to gravity $g$. There are many different ways to measure the value, but let's not bother with the details. Whatever the method, let's imagine that we have measured **5** values for $g$.



Lab Notebook

| | | |
|---|---|---|
| Measured values of accel. due to gravity: | 9.45 m/sec$^2$ | 9.26 m/sec$^2$ |
| | 10.04 m/sec$^2$ | 9.53 m/sec$^2$ |
| | 9.93 m/sec$^2$ | |

Ultimately, the question will be: What value do you report and what is the error associated with the value? Mathcad can help summarize the data, but we first have to get the numbers into Mathcad.

## Entering Data

Data is entered into a **table** in Mathcad. As an example, a table of time values is shown belo

$$time := \begin{bmatrix} 0.62 \\ 0.03 \\ 13.89 \\ 62.91 \\ 9.38 \\ 14.06 \end{bmatrix} \mathbf{s}$$

With the data labeled this way, we can ask Mathcad for the value of a specific number in the list by referring to the index. For example:

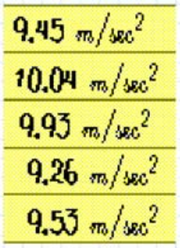
$time_1 = 0.03\ \mathbf{s}$ or $time_5 = 14.06\ \mathbf{s}$

When you use a range variable as an index, the steps must always be whole numbers. The variable $n$ is most commonly used for an index, but i,j, and $k$ are also conventional indices.

A table of data uses an index as a **numerical** subscript (for example: $time_i$ ). These subscripts are **not** the **literal** subscripts used to name variables ($x_o$ or $E_{total}$). Instead of being part of the name, numerical subscripts actually give Mathcad information (the order of the data points). The numerical subscript is entered by typing the left square bracket ( **[** ). There is also a slight difference in appearance:

| | | |
|---|---|---|
| **Literal** subscript: | type $t$ **[ctrl+-]** $n$ | $t_n$ |
| **Numerical** subscript: | type $t$ **[** $n$ | $t_n$ |

For practice, let's enter the accelerations due to gravity into a table. To avoid confusion with the predefined $g$, use the variable $G$. There is a trick for including units after the data entry, so you will just enter the numbers and leave the units for later.



| |
|---|
| 9.45 m/sec² |
| 10.04 m/sec² |
| 9.93 m/sec² |
| 9.26 m/sec² |
| 9.53 m/sec² |

$$\begin{bmatrix} a \\ \blacksquare \end{bmatrix}$$

In the open space, type:

**G   :**

After the equal sign, insert a matrix with **Ctrl + M**. Type in the other numbers (no units), separating each value with **Shift + Enter**.

$$G:=\begin{bmatrix} 9.45 \\ 10.04 \\ 9.93 \\ 9.26 \\ 9.53 \end{bmatrix}$$

If you succeeded in entering the data, then the list will be displayed to the right, when you type G[n = You have defined the variable $G$ to be a list of $n$ values. There are two ways of referring to tables:

**1**. Using $G_n$ refers to the individual numbers in sequence ($G_1, G_2 .. G_n$).

**2**. Using $G$ refers to the whole list, all at once.

Type $G_2$ =   $G_2 = 9.93$

Type $G$ =   $G=\begin{bmatrix} 9.45 \\ 10.04 \\ 9.93 \\ 9.26 \\ 9.53 \end{bmatrix}$

The distinction between a variable for a list ($G$) and one with a numerical subscript ($G_n$) is important and will become more clear as we proceed.

Here's the clever trick for giving a table of data the proper units after entering just the numbers:

**Trick**:   $G:=G\cdot\frac{\mathbf{m}}{\mathbf{s}^2}$   The result:   $G=\begin{bmatrix} 9.45 \\ 10.04 \\ 9.93 \\ 9.26 \\ 9.53 \end{bmatrix}\frac{\mathbf{m}}{\mathbf{s}^2}$

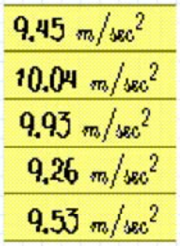
Notice that we redefined the whole list ($G$) to be the same variable, but with units of acceleration. Instead of having to type in the units over and over as you enter data, you can redefine the list to include the units afterward.

## The Mean

Notice that our values for the acceleration due to gravity are all different. How would you report a result for $g$?

Picking the value that is closest to the accepted value for $g$ would be **biased**. A better way would be to calculate the **mean** or average of your data. If the reason for the variation in the data is due to random errors, then calculating the mean will help to average out the fluctuations in the data and provide the best estimate of the true value of the quantity being measured.

How do you calculate the mean of our **5** accelerations?



| |
|---|
| 9.45 m/sec² |
| 10.04 m/sec² |
| 9.93 m/sec² |
| 9.26 m/sec² |
| 9.53 m/sec² |

$$\frac{9.45+10.04+9.93+9.26+9.53}{5}=9.64$$

True, you can add up all the values and divide by the number of values. Mathcad has a built-in function mean( ) for taking the mean of a list of numbers:

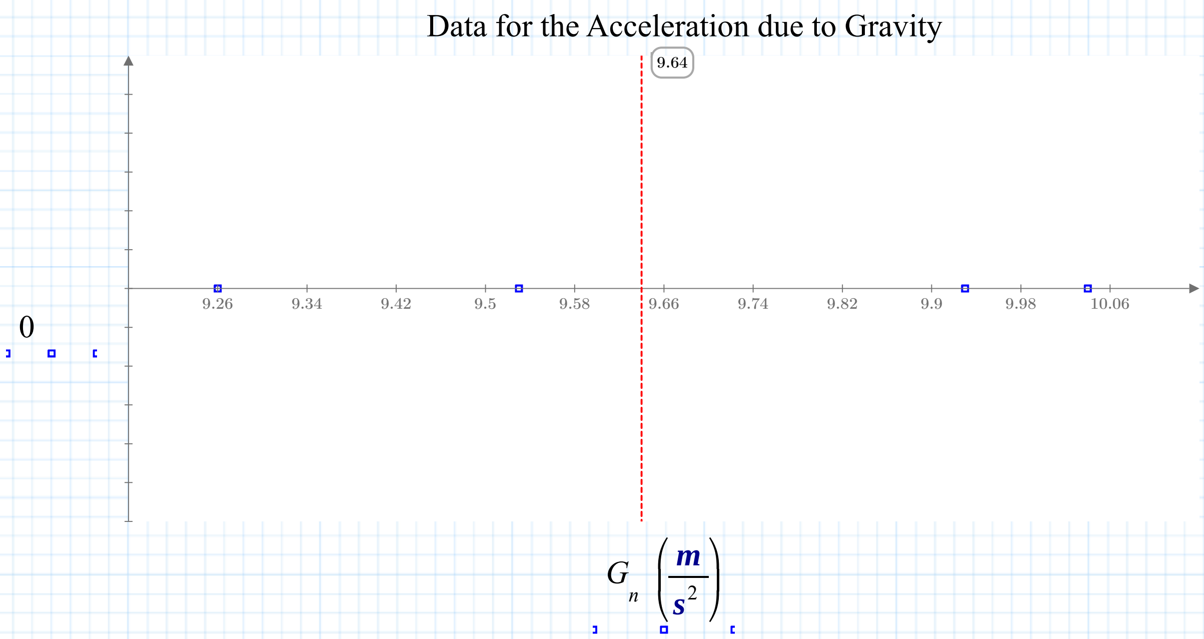
$$\text{mean}(G)=9.64\ \frac{\boldsymbol{m}}{\boldsymbol{s}^2}$$

**Notice** that we dropped the subscript! There is no such thing as an average for each number individually ($G_n$).

The argument of the mean( ) function is the list as a whole ($G$).

Here is a graph of your data with a marker at the mean value:

$n := 1, 2 .. 5$

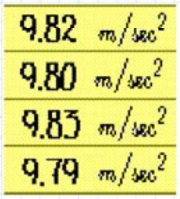


We will discuss how to graph data in Section 6.2, where we will discuss two sets of data that are related. In the graph above, we have simply graphed an arbitrary number (zero) against each of the data values $G_n$ .

## Standard Deviation

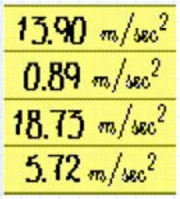
We now have a way of choosing a value of a measured quantity to report (the **mean**), but what is the error in the mean? We will answer that question below, but first we must discuss a concept known as the **standard deviation**. Let's look at two sets of experimental data acquired by two different students:

**Joe Science**



| | |
|---|---|
| 9.82 | $m/sec^2$ |
| 9.80 | $m/sec^2$ |
| 9.83 | $m/sec^2$ |
| 9.79 | $m/sec^2$ |

**Joe Thumbs**



| | |
|---|---|
| 13.90 | $m/sec^2$ |
| 0.89 | $m/sec^2$ |
| 18.73 | $m/sec^2$ |
| 5.72 | $m/sec^2$ |

Notice that Joe Science has very consistent results, while Joe Thumbs' values vary all over the place. Now look at their mean values:

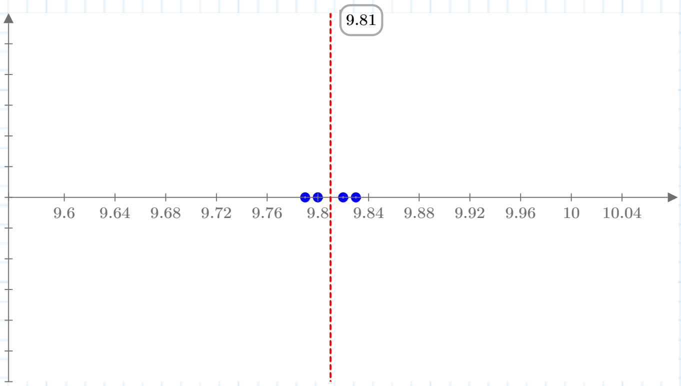
**Joe Science**

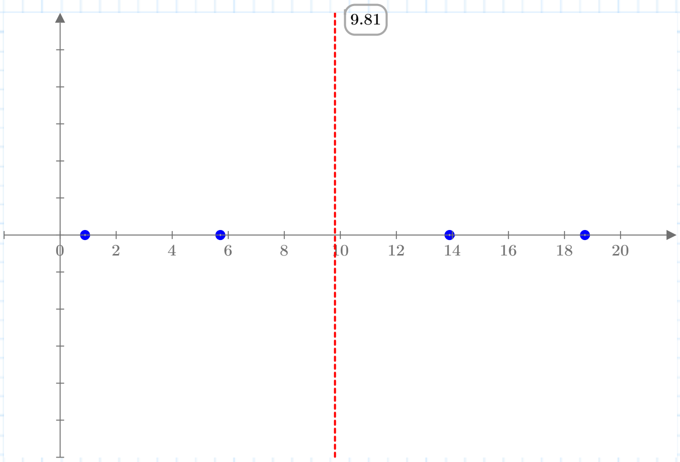
$$\frac{9.82+9.80+9.83+9.79}{4}=9.81$$

**Joe Thumbs**

$$\frac{5.72+0.89+18.73+13.9}{4}=9.81$$

Even though both students would report the same mean, Joe Science's' results are more reliable than Joe Thumbs because on average Joe Science's' values are closer to the mean that he calculates. You can see the difference on the graphs below (notice the difference in scale between the two graphs).





This example shows that simply concentrating on the mean can be deceiving. To evaluate the reliability of the result, we must also have a way to express the amount of **variation** in the data. Are the values tightly grouped around the mean value or broadly spread out? This variation will ultimately be related to the error in the mean.

Let's calculate how far each student's values are away from their mean value: 9.81

| **Joe Science** | **Joe Thumbs** |
|---|---|
| $9.81-9.82=-10\cdot10^{-3}$ | $9.81-5.72=4.09$ |
| $9.81-9.80=10\cdot10^{-3}$ | $9.81-0.89=8.92$ |
| $9.81-9.83=-0.02$ | $9.81-18.73=-8.92$ |
| $9.81-9.79=0.02$ | $9.81-13.9=-4.09$ |

You can see that Joe Thumbs always measured values further from the mean than did Joe Science. Assuming that the variations in data points are mostly due to mistakes made during the experiment, we will refers to the variations as **error** in the data. Joe Thumbs likely used a very sloppy lab technique or very imprecise equipment in collecting his data.

We must first calculate the **average amount of error in each data point**, which is called the **standard deviation** (*standard* meaning average). Unfortunately, we cannot simply take the mean of the deviations or the positive and negative deviations will just cancel one another and we will get zero.

To avoid this problem the **standard deviation** uses the squares of each deviation so that the numbers are all positive (for example, $(-0.02)^2=0.0004$ ) and then takes the average. The square root of this average value is then calculated to make up for having squared each value individually. The general formula for the standard deviation is written in “summation notation” as

$$Standard\_deviation = \sqrt{\frac{1}{N}\cdot\sum_{n=0}^{N}\left(x_n - mean(x)\right)^2}$$

where $N$ is the total number of data points and $x_n$ refers to each data point individually, while $mean(x)$ is the average value of the whole data set $x$ .

Mathcad has a function stdev( ) that calculates the standard deviation for a set of data. The standard deviation is an estimate of the random errors that occurred during the measurements (assuming that the physical thing we are measuring does not vary!).

Use the stdev( ) function to calculate the standard deviation for your set of data $G$. As with the mean, you must use the data set as a whole ( $G$ not $G_n$ ).

$$\text{stdev}(G)=0.3\ \frac{m}{s^2}$$

What does that number say about your set of data?

## Error in the Mean

Knowing the standard deviation associated with a data set, we can calculate the error in the mean, as follows. We define the error in the mean with the following user-defined function:

$$errmean(data, N) := \frac{\mathrm{stdev}(data)}{\sqrt{N}}$$

In this equation $N$ is the number of data points (for example, and $data$ refers to a particular list of data points.

**Important: It is the error in the mean that is reported as the error associated with the value of your measured quantity.** In words, the error in the mean is the average error in each data point (the standard deviation) divided by the square root of the number of data points.

For example, in Joe Science's experiment the standard deviation associated with his data set is $0.02 \cdot m \cdot s^{-2}$. Therefore, since $N = 5$, the error in the mean of $g$ is

$$\frac{0.02 \cdot \mathbf{m} \cdot \mathbf{s}^{-2}}{\sqrt{5}} = \left(9 \cdot 10^{-3}\right) \frac{\mathbf{m}}{\mathbf{s}^2}$$

Therefore Joe Science's value of $g$ is g = 9.81 ± 0.009 $m \cdot s^{-1}$. Joe Thumbs' data has a standard deviation of “6.94m”$\cdot sec^{-1}$. Consequently, his error in the mean is

$$\frac{6.94 \cdot \mathbf{m} \cdot \mathbf{s}^{-2}}{\sqrt{5}} = 3.1 \frac{\mathbf{m}}{\mathbf{s}^2}$$

As expected, this value is much larger than Joe Science's error. Joe Thumbs would report his value of $g$ as g = 9.8 ± 3.1 $m \cdot s^{-1}$.

Use the user-defined function $errmean(■, ■)$ to calculate the error in the mean for your set of data $G$.

$$errmean(G, 5) = 0.13 \frac{\mathbf{m}}{\mathbf{s}^2}$$

## Fractional Error

To get an impression of whether the standard deviation associated with a set of data is “big” or “small,” you must compare the deviation to the mean value. For example, making errors of a few meters is still very accurate when you are talking about the size of the earth, but is very poor when discussing the size of an atom.

The ratio of the standard deviation to the mean value is defined as the **fractional error**. Below, we have calculated the mean, standard deviation, and fractional error for the data sets taken by the two students.

| | **Joe Science** | **Joe Thumbs** |
|---|---|---|
| mean: | $9.81 \cdot \boldsymbol{m} \cdot \boldsymbol{s}^{-2}$ | $9.81 \cdot \boldsymbol{m} \cdot \boldsymbol{s}^{-2}$ |
| standard deviation: | $0.02 \cdot \boldsymbol{m} \cdot \boldsymbol{s}^{-2}$ | $6.94 \cdot \boldsymbol{m} \cdot \boldsymbol{s}^{-2}$ |
| **fractional error**: | $\frac{0.02}{9.81} = 0.2\ 1\%$ | $\frac{6.94}{9.81} = 70.74\ 1\%$ |

Notice that the fractional error immediately gives you a feeling for how large or small the standard deviation is in comparison to the mean value. While both students found the same mean acceleration, you can see from the fractional error that the second student is “all thumbs.”

Compute fractional error for your data set $G$ (the ratio of the $stdev(G)$ to the $mean(G)$ value).

**Hint:** Just use the functions in the expression.

Insert the $\%$ in the units placeholder.

$$\frac{\text{stdev}(G)}{\text{mean}(G)} = 3.06\ 1\%$$

## Practice

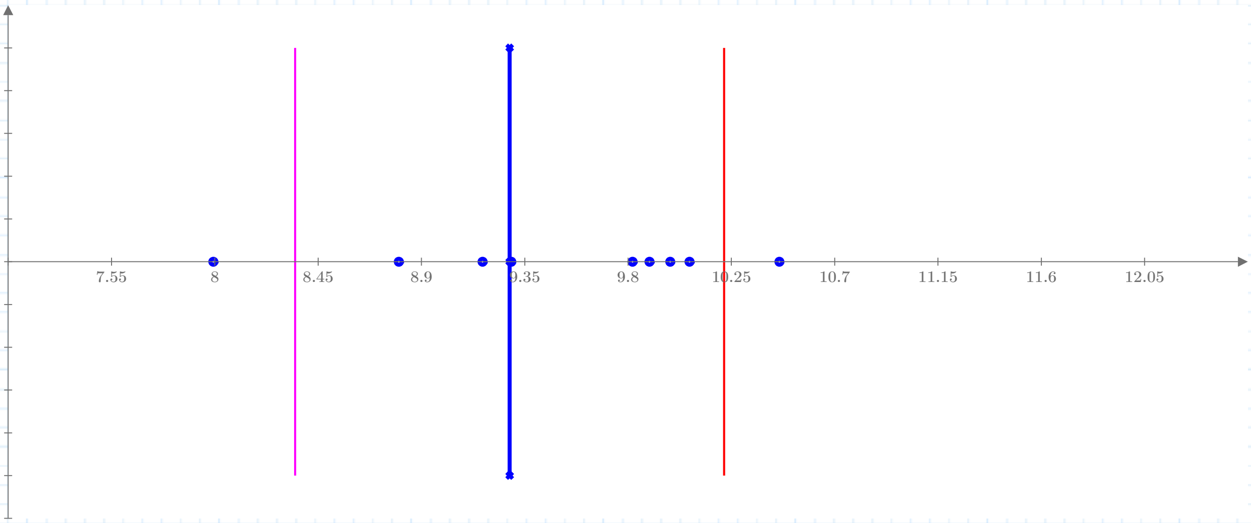
To become more familiar with these new ideas, we have created a way for you to conduct multiple experiments and observe the **mean, error in the mean,** and **fractional error**. The **10** data points below (called *data* ) are actually the result of a random number generator. To get new values, just click on the equation below, and press [F9].
Then keeping your cursor on the data points, scroll down until you can see the graph and values below it.

$$data := \left( \left( 9.81 + \mathrm{rnorm}\left(10, 0, \mathrm{rnd}(1)\right)\right) \cdot \frac{\boldsymbol{m}}{\boldsymbol{s}^2} \right)$$

Index for data: $k := 1 \,..\, 9$

Press [F9] repeatedly to see how different data numbers affect the mean, error in the mean, and fractional error listed below the graph.

$$data_k = \begin{bmatrix} 7.99 \\ 9.98 \\ 9.29 \\ 8.8 \\ 9.89 \\ 10.46 \\ 10.07 \\ 9.82 \\ 9.17 \end{bmatrix} \frac{\boldsymbol{m}}{\boldsymbol{s}^2}$$



### Mean

$$\text{mean}(data) = 9.28\ m \cdot \mathbf{s}^{-2}$$

### Error in the Mean

$$\frac{\text{stdev}(data)}{\sqrt{10}} = 0.3\ m \cdot \mathbf{s}^{-2}$$

### Fractional Error

$$\frac{\text{stdev}(data)}{\text{mean}(data)} = 10.06\ 1\%$$

This **global** variable definition creates a line (from $-1$ to $1$) for the standard deviations on the last graph:$f$

$$\gamma \equiv -1 \,..\, 1$$

The mean, standard deviation, and the related quantities fractional error and error in the mean are known as **descriptive statistics** that are used to summarize a set of data. The underlying assumption is that you are repeatedly making the same measurement of a relatively stable physical quantity. If that is the case, then these statistics are invaluable for describing a set of data.

The exercises in the next section introduces you to the analysis of two sets of data that are related in some way (for example, the velocity of an object and its position). You will also learn how to graph data so that you can understand the relationship and propose a theoretical explanation.