How Many Data Points are Enough?

Lilienthal's rule: *If you want to fit a straight-line to your data, be certain to collect only two data points. A straight line can always be made to fit through two data points.*

Corollary: *If you are not concerned with random error in your data collection process, just collect three data points.*

Corollary: *If you can make measurements completely without error, you need collect only one set of data and never need to repeat any measurements.*

Clearly these are wrong ways to collect and interpret data – for a line, and indeed any curve, can be made to pass through just two points. Random errors in data collection can also cause problems with the interpretation of graphical data if the number of data points is too few. More data will allow for the averaging out of random error.

How much data, then, should be collected before generating a graph? That's a good question, and one that is rarely addressed adequately in science courses that have laboratory components. Perhaps it is best to provide some examples of inadequate data collection to show the problems that arise from using an inadequate number – and sometimes more importantly an inadequate range – of data points.

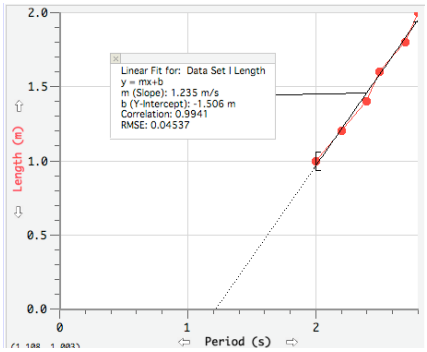
Inadequate number of data points: Consider a task in which a student is asked to determine the relationship between Fahrenheit (F) and Celsius (C) temperature scales using a dual-scale thermometer as the source of data. If only three data points are collected, then a best-fit equation will be strongly biased by errors in the data. For instance, the student generates a table of C versus F temperatures from the inspection of the C/F thermometer. The student notes that 0^{o}C corresponds precisely to 32^{o}F and that 212^{o}F corresponds precisely to 100^{o}C. The student then uses the ambient air temperature to see that 73^{o}F corresponds to 23^{o}C approximately (actually $22.777\ldots^{o}$C). A best-fit regression equation will result in a non-linear relationship due to this reading error. If multiple data points are collected (say 6), then the errors tend to average themselves out and a linear relationship would be clearly evident. Hence, as a general rule, the more data points the better – up to a point. Too many data points constitute needless effort and a waste of time.

It is often difficult to decide just how much data to collect, but here are some simple guidelines: if the relationship is expected to be linear, then fewer data points are necessary to arrive at the proper regression formula; if the relationship is expected to be non-linear, then more data should be collected in the area where the curve is expected to rapidly change, and less data may be collected where the curve is expected to be more linear. For instance, in the relationship between the period of a pendulum and its length, more data should be collected at shorter pendulum lengths and less data should be collected at longer pendulum lengths. Consider the following example.

Inadequate range of data: Say that someone is attempting to determine the period a simple pendulum – a weight suspended at the end of a string. Length-versus-period data are collected

for a variation of length of only 1 meter starting at 1 meter and going out to 2 meters. The following data are collected and graph generated.

Length (m)	Period (s)
1	2.0
1.2	2.2
1.4	2.4
1.6	2.5
1.8	2.7
2.0	2.8



Because of slight random errors in data collection (which are to be expected), it at first appears that the data best fit a linear relationship. Even if a physical model (derived from **dimensional analysis**) is suggested (e.g., the best fit relationship must pass through the origin), it is difficult to reconcile the present data that clearly appear not to pass through the origin. Question: What is the problem? Answer: Not a wide enough range of data has been collected. Data should be collected in a concentrated fashion for shorter lengths of the pendulum, whereas fewer data points need to be collected at greater lengths of the pendulum. *As a general rule, the maximum value of the independent variable should be at least 5 (preferable 10) times the minimum value of the independent variable used in making a graph.*

Dimensional analysis will show that the expected period-length relationship for the simple pendulum is parabolic. If data are collected too far out along the parabola and not enough close in, then that limited set of data clearly suggests a straight-line relationship. These data are reasonably correct for the interval of length studied however. The problem associated with generating a more general solution of the period-length relationship lies in not collecting a wide enough range of data to represent the phenomenon accurately over a wide range. For this reason, *extrapolation is not generally merited beyond the range of the data.*

Inadequate separation of data points: Consider another variation of this problem – inadequate separation of data points. A student attempts to determine the relationship between the period of a torsion pendulum and its moment of inertia. The student varies the moment of inertia in small increments and measures the resulting period. This results in the following data set:

Moment of Inertia ($kg\,m^2$)	Period (s)
.060	.15
.061	.17
.062	.16
.063	.18

Note the lack of a trend in the data. While the moment of inertia appears to go up uniformly, the period is more or less random. This apparent randomness results from small measurement errors that, while they may be small in absolute terms, are large in relative terms. (See **relative error**

and **absolute error**.) The problem can be overcome by collecting data points that are separated to a much greater degree. That is, the student should attempt to vary to moment of inertia of the torsion pendulum by increasing the moment of inertial not in tiny steps, but in much larger steps.

Errors in data collection: **Repeated measures** will help to minimize the amount of error in an observation and of relationships derived from graphs associated with those data. All well-done human measurements include some randomized error. Such error results from the limited ability of humans to make measurements. Using a stopwatch, for instance, results in data that can easily be "off" 0.2 seconds or more due to the lack of instantaneous hand-eye coordination. Besides, knowing just when to start and stop a stopwatch can be variously interpreted. Even the use of precise scientific instruments will result in measurements that contain errors, albeit smaller than those conducted with human intervention.

Most data collection that students will do in a simple laboratory setting might be fruitfully completed using a number of observations of the same phenomenon, and then taking the average. An average will help to find the central tendency of a series of measurement around the central value and will result in a more precise indication of the actual value of a repeatable phenomenon such as the period of a swinging pendulum.

To what extent students make repeated measures will depend on a number of factors such as the use of precise scientific instruments, the degree of human interaction, the amounts of relative and absolute errors, how low it takes to repeat an observation, the degree of precision needed, time allocated for a particular laboratory activity, and so on. In a classroom situation the teacher is best suited to making this judgment.

A general rule for repeated measures: *If the precision of the measurement is high, the number of repeated measurements can be low; if the precision of the measurement is low, the number of repeated measurements should be high.*

So, as long as no human blunders are made, an acoustical motion detected can be used once for measuring the motion of an object. The use of a stopwatch, on the other hand, is subject to substantial error and repeated measures should be made if assessing motion. The average of the repeated measures is then reported and used for the development of a graph if appropriate.

How many data points are enough?: As an answer to the initial question, a simple and fast rule for introductory labs would be to collect 6 data points minimum. This is, admittedly, an arbitrary number but one that seems to work well in the introductory setting. The selection of the data points must take into consideration the following:

- If the accuracy of the measurements is low, each data point used to make a graph should consist of a number of observations that have been averaged.
- The maximum value of the independent variable should be at least 5 time (10 times is better) its minimum value.
- If the function is expected to change rapidly, then data collection should concentrate in those regions of a graph where the data are expected to vary most quickly.