

# Sequential Introduction of Data Analysis Methods in the Modern Lab

Timothy Roach

*Department of Physics, College of the Holy Cross, 1 College St., Worcester, MA 01610*

**Abstract:** A major goal of many intermediate physics labs is learning methods of data analysis. In our Modern Lab course we introduce these methods in a planned sequence, with labs explicitly designed to match the sequence, so that students learn increasingly more sophisticated methods as the semester progresses. The first lab has them investigate repeated measurements of a single quantity (the speed of electromagnetic pulses and speed of light) and introduces the concept of error propagation. In the second lab they use a functional relation ( $\lambda$  vs.  $\sin \theta$ ), for calibration of a diffraction grating, using residuals to optimize the fit. Later labs introduce Gaussian and Poisson probability distributions, and Least-Squares fitting of functions (including non-linear minimization). In addition, we provide here a few examples of how either methods or experiments can be adapted in order to support a coherent sequence of learning.

**Keywords:** data analysis methods, physics teaching labs.

**PACS:** 01.50.Qb, 07.05.Hd

## INTRODUCTION

Learning data analysis methods in the course of an advanced physics lab is often challenging for the students because they are simultaneously learning several other new topics or skills, including the physical theories underlying the experiments, how to use scientific equipment, and how to write about or present their findings. At the same time, teaching data analysis methods can be challenging for a lab instructor because different student groups in a class will often be doing different experiments, making it difficult to take a consistent class approach to data analysis.

As in learning any topic, students should have opportunities to practice a method multiple times, in different contexts. The class should not introduce too many new methods or new quantities in one class; these must be sufficiently experienced before building on them to higher concepts or moving to a different strand.

There are usually more than one choice of analysis method to extract information from experimental data, and the best method for research may not be the best method for student learning at a particular point in their education. With some planning, methods can be chosen for a sequence of experiments to build a logical sequence in which data analysis methods are introduced. The sequence of experiments can itself be configured to support such a logical sequence. Providing opportunities for students to repeatedly apply earlier methods in new contexts is also a key to learning that will last.

## SEQUENCE OF DATA ANALYSIS METHODS

In our Modern Physics Lab course, we introduce major data analysis methods in this approximate sequence: measuring a single quantity (error estimation, error propagation, combining errors from different sources); probability distributions and statistical methods (Gaussian and rectangular distributions, mean and standard error, weighted average); measuring and modeling a functionally dependent quantity<sup>1</sup> (fitting data to a model, residuals, least-squares criterion); and further probability distributions (Poisson, Gaussian, exponential). Below we describe how these are implemented.

We have not made any attempt to quantitatively assess student learning gains in comparison to other approaches. Our experience is that it is more straightforward to *teach* data analysis using this sequence compared to what we did before, and there seems to be less confusion among the students about how and when to make use of the various analysis methods.

### Measuring a Single Quantity

Measuring a single quantity is a logical and simple place to begin, but measuring a primary quantity (for example, a distance measured with a meter stick or a weight measured with a scale) is not very interesting. In the first session of our course, students measure the speed of electromagnetic pulses in coaxial cable,

calculated from the two primary measurements of cable length and travel time. The apparatus (high speed oscilloscope and coaxial cables) and related concepts (EM waves traveling in an inductive capacitive medium) are usually quite novel to the students, yet the mathematics are simple enough to allow them to readily take in the methods of error propagation.

Specifically, they measure the time ( $T$ ) for a pulse to travel to the end and back of a cable of measured length ( $L$ ), from which, the speed is determined as  $v = 2L/T$ . Estimated errors in the primary quantities we designate by  $\alpha_L$  and  $\alpha_T$ . Error propagation (using the usual linearization approximation, for small errors), produces uncertainties  $\alpha_v^L = \left| \frac{dv}{dL} \right| \alpha_L = \frac{v}{L} \alpha_L$  and  $\alpha_v^T = \left| \frac{dv}{dT} \right| \alpha_T = \frac{v}{T} \alpha_T$ , where  $\alpha_v^L$  and  $\alpha_v^T$  are the partial uncertainties due to  $L$  and  $T$ , respectively. They compute the total uncertainty as  $\alpha_v = \sqrt{(\alpha_v^L)^2 + (\alpha_v^T)^2}$ .

The size of the fractional errors from distance ( $\alpha_L/L$ ) and time ( $\alpha_T/T$ ) are quite different, providing an example of the relative importance of error contributions. Note that these methods do not require knowledge of probability distributions, except for the combining of errors in quadrature, which is strictly justified only for errors from Gaussian distributions.

This learning is reinforced the following week, by applying the same methods in a slightly different experiment, measuring the speed of light. A pulse of light from a diode laser is reflected from a mirror at near-normal incidence, back to a photodetector. The mathematics is nearly identical to the previous week, but incorporates subtraction out of a reference point ( $L_0, T_0$ ), which eliminates poorly known quantities such as the response times of the laser and photodetector. Because the speed of light is a known quantity, students can determine the actual error in their measurement and compare to their error estimates.

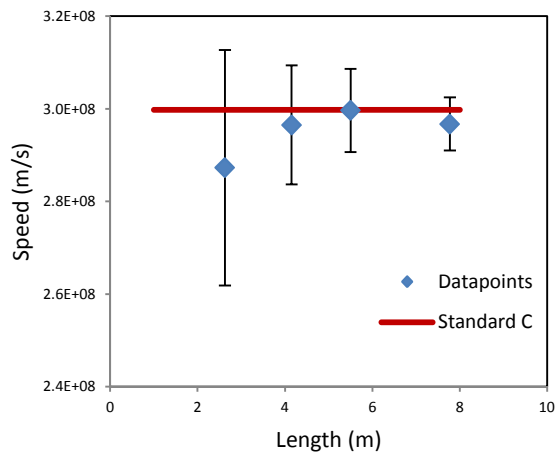
Investigation of multiple trials is also done with this data; this is discussed below.

## Probability Distributions and Statistical Measures

During the second lab session (Speed of Light experiment), probability distributions are introduced, first as a tool for quantifying what we mean by an error estimate of a primary measurement. For example, the instrumental resolution of time measurement for our digital oscilloscopes is 0.4 ns, representing an uncertainty with a rectangular probability distribution  $\pm 0.2$  ns. On the other hand, if the reflected laser signal

is weak, repeated measurements will vary, fluctuating such that "much of the time" they fall within (for a given case)  $\pm 1$  ns, which we model as a Gaussian distribution with  $\sigma = 1$  ns, and equate "much of the time" with 68%. This is also a time that the GUM (*Guide to the expression of uncertainty in measurement*<sup>2</sup>) concepts of Type A and Type B uncertainties can be introduced, if desired.

Repeated measurements of a single quantity allow one to apply basic statistical measures such as mean, standard deviation, and standard error. For the speed of light lab, the repeated measurements can be a set of trials, one from each group. (This is useful only if the class is large.) Alternatively, each group can produce a set of trials taken at different distances. A useful aspect of this is that it makes clear (after use of error propagation) the value of longer distances and times for improving the precision of the measurement. For example, Fig. 1 shows the decrease in uncertainty (error bars) with increase in travel distance, a result primarily due to decrease in the relative uncertainty in time measurements.



**Figure 1:** Repeated trials of a single quantity measurement. Here, the speed of light measurements show improved precision at longer distances.

A disadvantage of using trials at different distances is the one cannot properly combine them using a simple mean. Instead, we have the students compute a weighted mean from their values and a weighted standard error.

## Curve Fitting

We progress from single quantities to functional relations, that is to say, curve fitting. While there are well-known matrix methods<sup>3</sup> and readily available computer tools to automatically fit functions to data, we avoid these, at least initially, because we want to

build student understanding of the principles underlying the method. First we emphasize the idea that there is a mathematical model, in most cases derived from other physical laws, that we expect should describe the data. Next we have the students graph the model equation (having adjustable parameters) on the same plot as their data points, and observe how the model changes as parameters are adjusted. We use Excel, but this can also be done in Mathematica, Origin, or other math or data graphing software. They optimize the model first by eye. Then we have them plot the residuals (deviations between the model and data), which aids their parameter optimization and can also provide insight into the random or systematic character of deviations.

Only then do we introduce the least-squares criterion. The sum of the squared residuals ( $SSR$ ) is readily calculated in the spreadsheet, and students improve their fit by further manually adjusting the parameters to minimize the  $SSR$ . We later show them computational tools to do this more quickly, such as the Solver optimization tool in Excel. They compare their optimal results to other quick tools, for example creating a Trendline, which is of course, based on the same least-squares criterion.

In our most recent incarnation of our lab course, students first learn curve-fitting by applying a straight line model to calibration of a diffraction grating, using known Hg wavelengths (for a Balmer spectroscopy lab). With the expected model  $m\lambda = d \sin \theta$ , we have them plot  $m\lambda$  versus  $\sin \theta$ , with one adjustable parameter,  $d$ .

This same general approach for curve fitting is applied later in the course, at least twice, usually to models that are not a straight line. Examples of non-straight lines include  $\lambda$  vs.  $n$  for Balmer wavelengths,  $f$  vs.  $L$  for resonant frequencies of hollow tubes, and  $R$  vs.  $t$  for activity of a short-lived radioactive isotope.

Estimating uncertainties in the model parameters we do only roughly, having students observing the fit and the residuals, to determine how far a parameter can be changed before the fit is no longer good, as judged by eye. More quantitative measures require more time than is available in a one-semester course, and require better knowledge of data point uncertainties than we usually have.

### More on Probability Distributions

Students delve more deeply into probability distributions primarily in our nuclear rate experiments, for which the Poisson distribution describes chances of specific counts being obtained. Using computer count acquisition, they can in reasonable time obtain

sufficient data to make a good histogram, to which theory can be compared. This is done both for a small average number of counts ( $\bar{N} \approx 1$ ) and a large number of counts ( $\bar{N} \approx 300$ ). In both cases they examine the expected relation  $\sigma = \sqrt{\bar{N}}$ . For the larger count situation, the Poisson distribution closely approaches a Gaussian, and this is used to explore the character of  $1\sigma$  and  $2\sigma$  confidence ranges.

## ADAPTING METHODS AND EXPERIMENTS

Building a coherent sequence of data analysis steps into a laboratory course often requires flexibility in either application of methods to a particular experiment and/or flexibility in ordering of experiments. For example, the data taken for the speed of light lab might more preferably have been analyzed by fitting the distance to a straight-line model as a function of time, with the slope (speed) as an adjustable parameter. This would be a cleaner scientific approach, because the detector and laser response times affect the offset but not the slope. However, we choose (in week two of our course) to arrange the data as multiple measurements of a single quantity, so that students can more fully explore this more basic approach. Extracting the speed from the slope can be done later in the course, if one wishes, and indeed we have sometimes used re-analysis of the speed of light data as the students' first introduction to curve-fitting.

Another example of adapting approaches is that we have sometimes moved part of our nuclear rate experiments to the first or second week of the semester, focused simply on the variability of repeated measurements of the counts in a fixed time, in order to explore the meaning of a probability distribution. No detailed knowledge of the radiation detector or decay processes is needed.

Often the lab instructor must choose experiments subject to significant constraints imposed by limited quantities of apparatus (so that several experiments are underway at the same time, by different student groups) or those imposed by a parallel lecture course (for example, putting gamma spectroscopy or particle experiments to the end of the semester because these topics are covered late in a modern physics lecture course). We have shown here examples of how one can adapt the analysis methods chosen for each experiment so as to provide a coherent, sequential introduction to some of the most important data analysis methods for students in an intermediate physics laboratory course.

## REFERENCES

---

1. Incorporating learning goals about modeling into an upper-division physics laboratory experiment, Benjamin Zwickl et al., *Am. J. Phys.* 82 (9), September 2014.
2. Evaluation of measurement data — Guide to the expression of uncertainty in measurement, JCGM 100:2008, BIPM.
3. See, for example, *Measurements and Their Uncertainties*, by Ifan Hughes and Thomas Hase, Oxford U. Press, 2010.