Introduction to Statistics and Data Analysis

RSI 2005 Staff

July 15, 2005

Variation and Statistics

- Good experimental technique often requires repeated measurements of the same quantity

- These repeatedly measured values will usually not be identical, and they may not match a theory or what you expected. You must explain why:
  - **Statistical Error** (σ)
  - **Systematic Error** (Δ)

Statistical Error (σ)

- Statistical error is due solely to random fluctuations in measuring process

- You can [and should!] carefully calculate σ

- You can [and should!] reduce σ by averaging over many independent, uncorrelated measurements

Central Limit Theorem: As the number N of independent, uncorrelated measurements of the same random variable approaches ∞, the average value of the sum is proportional to N while the standard deviation (σ) only grows as $\sqrt{N}$

Calculating Statistical Error

- **Discrete case** [counting]: $\sigma = \sqrt{N}$, so the total count, with uncertainty, is $N \pm \sqrt{N}$. If you're counting things in different bins, use this to find the error in each bin.

- **Continuous case** [e.g. measuring somebody's height]: make many repeated measurements, and use the following formula:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - m_x)^2}$$

where x_i is the measured value on the i^{th} trial and

$$m_x = \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$$

.

Systematic Error ($\triangle$)

- Systematic error describes deviations between data and theory that result from an incorrect theory, a faulty experimental apparatus, or incorrect data reduction.

- Should be eliminated if possible. Otherwise, it should be explained and estimated in your discussion of results.

- Thus, any reported value x should have the form

$$x \pm \sigma_x \pm \Delta_x$$

Reporting Data: Numbers

- **RULE #1:** Any reported value must be reported along with its uncertainty — both statistical and systematic. The statement "We found $\pi = 3.149$" is worse than incorrect, as we cannot evaluate its correctness. Stating "We found $\pi = 3.149 \pm 0.002$" is at least incorrect, in that we know the deviation from truth is due to something systematically wrong with the experiment or analysis.

- **RULE #2:** Any reported value must be accompanied by units. With the exception of tables in which units are specified in the top row, every number you put down in your paper should be dimensioned. Saying "We measured a lifetime of $(4.23 \pm 0.07) \times 10^5$" is completely meaningless.

Precision, Accuracy, and Significance

- **RULE #3:** Only use the word **precise** when talking about the statistical error of a measured quantity. A precise measurement has low statistical error.

- **RULE #4:** Only use the word **accurate** when discussing the systematic error of a measurement or process. An accurate measurement returns a value that is very close to the "true" value.

- **RULE #5:** Only use the word **significant** when discussing results that can agree or disagree with a hypothesis given some error probability threshold. Always specify this threshold when you use the word **significant**.

Error Propagation

- Suppose you measure some quantity X and obtain a value of $x \pm \sigma_x$, but you want to know the value of $Y = X^2$.

- Clearly the average value of Y you would report is $y = x^2$, but what is σ_y?

$$\sigma_y^2 = \sigma_x^2 \left[\frac{dY}{dX}\right]^2_{X=x}$$

- For more variables, errors add in quadrature: If $Y = f(U, V)$, and if errors in U and V are uncorrelated, then

$$\sigma_y^2 = \sigma_u^2 \left[\frac{\partial Y}{\partial U}\right]^2_{U=u} + \sigma_v^2 \left[\frac{\partial Y}{\partial V}\right]^2_{V=v}$$

Error Propagation Examples

- Suppose $Z = X \pm Y$. Then $\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}$.

- Suppose $Z = XY$ or $Z = X/Y$. Then $\sigma_z = z\sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$.

- Suppose $Z = a^{bX}$. Then $\sigma_z = zb\sigma_x \ln a$.

Raw Data is Often Uninteresting

- May depend on factors specific to your experiment
- Often involves "uninteresting" quantities [e.g. counts per bin, intermembrane glucose concentration]
- Want to convert the data into more interesting physical quantities that can be compared with theoretical models and other experiments. [known as **data reduction**]
- Necessary for both quantitative and qualitative measurements

Plotting Data

- Always label your axes with quantity being measured (e.g. height) and units (nm)

- You MUST include [usually vertical] error bars to show uncertainty in the dependent quantity being measured. This applies to bar graphs, scatter plots, and almost any other graph you will use. Remember that for counting measurements, this error is just $\pm\sqrt{N}$.

- If there is uncertainty in any independent quantity, you must indicate this with error bars as well [usually horizontal]. If the uncertainty in the dependent quantity is correlated to the uncertainty in the independent quantity, you should plot an error ellipse.

Fitting Curves to Data I

- Goal of is to find a function that can be used to describe your data

- Why bother fitting? It allows us to:
  - extract interesting parameters from our data
  - compare our data to a model or theoretical prediction
  - interpolate or extrapolate to make predictions

- Don't ever just "connect the dots" with plotted data. Doing so would suggest the existence of data or theories that aren't there.

Fitting Curves to Data II

- To fit, first choose a functional form for the fit with designated parameters that can be varied.

- Vary parameters to globally minimize a given error function. [Do this with a computer and fitting software, not by hand!]

- The most common function used for error is the chi-squared (χ^2) metric. Other metrics are used to weigh errors in different ways.

Minimizing χ^2

- The χ^2 metric is given by

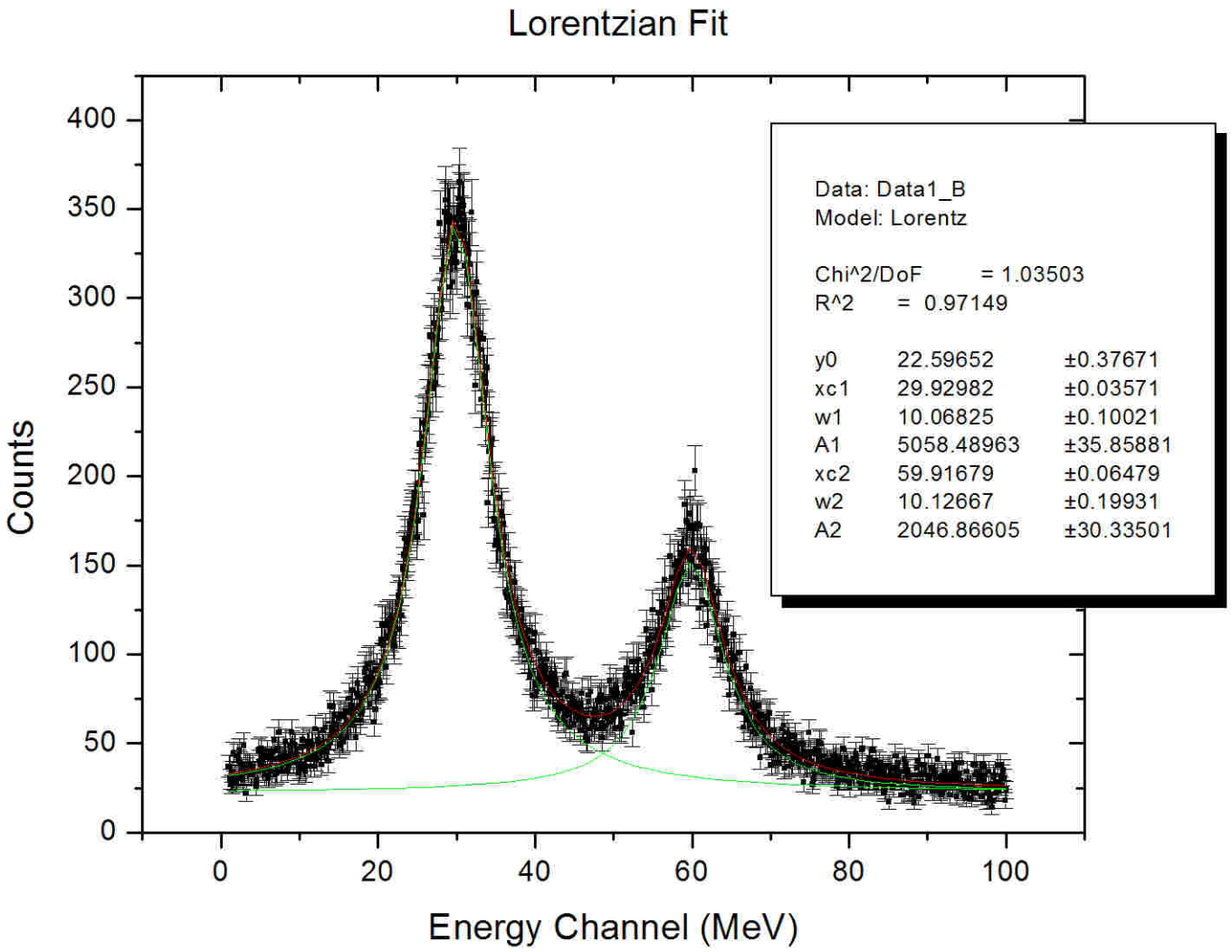
$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \widehat{x}_i)^2}{\sigma_i^2}$$

  Here, x_i are your measured data, $\widehat{x}_i$ are the predictions of your function [these values change as you vary parameters], and σ_i is the uncertainty in x_i.

- Finding a global minimum in the χ^2 metric amounts to minimizing the sum of squared differences between your data and the model.

- You want $\chi^2/DoF \approx 1$, where DoF is the degrees of freedom in your fit [number of data points - number of free parameters in model]

Example


Lorentzian Fit
Counts
Energy Channel (MeV)
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100
Data: Data1_B
Model: Lorentz
Chi^2/DoF = 1.03503
R^2 = 0.97149
y0 22.59652 ±0.37671
xc1 29.92982 ±0.03571
w1 10.06825 ±0.10021
A1 5058.48963 ±35.85881
xc2 59.91679 ±0.06479
w2 10.12667 ±0.19931
A2 2046.86605 ±30.33501