Statistics, Curve Fitting, and Parameter Estimation in 8.13

A Practical Approach

Javier M. G. Duarte

Department of Physics
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

October 8, 2010

1 Statistics

2 Probability Density Function, Histograms, & Error Bars

3 χ^2 Test Statistic

4 Minimization Techniques

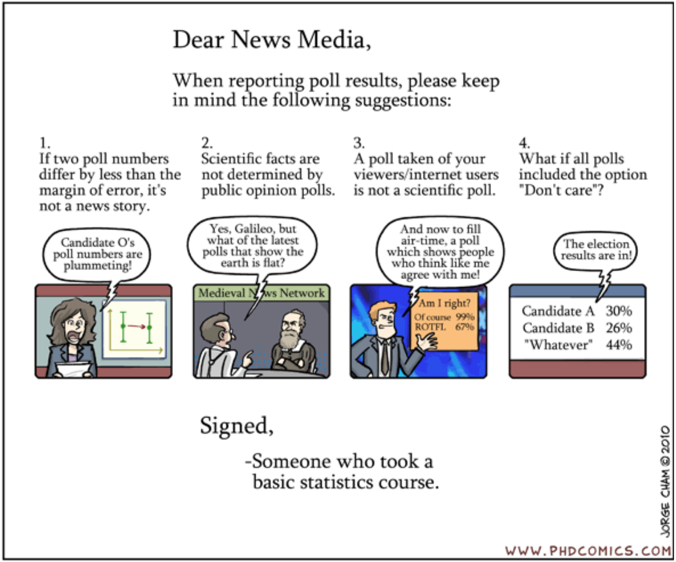
5 Tips & Tricks

6 References

Statistics: Who cares?

- We (as aspiring physicists) do!
- We employ statistics (study of large systems) as a way of making sense of fluctuations in our world (data)
- We use well-defined techniques to quantify measurement, and uncertainty
- Every random variable has some distribution (in theory)
- If we know what the distribution should be, we can peform a *FIT* to the data using the expected curve
- $\Rightarrow$ Extract some important, physical parameters (mean, σ, decay time constant, etc)
- Most of the time, (Central Limit Theorem) it's a Gaussian
- If you're counting something, it's probably a Poisson[1]

[1] And Poisson $\stackrel{\mu\to\infty}{\longrightarrow}$ Gaussian


Dear News Media,
When reporting poll results, please keep in mind the following suggestions:
1. If two poll numbers differ by less than the margin of error, it's not a news story.
Candidate O's poll numbers are plummeting!
2. Scientific facts are not determined by public opinion polls.
Yes, Galileo, but what of the latest polls that show the earth is flat?
Medieval News Network
3. A poll taken of your viewers/internet users is not a scientific poll.
And now to fill air-time, a poll which shows people who think like me agree with me!
Am I right?
Of course 99%
ROTFL 67%
4. What if all polls included the option "Don't care"?
The election results are in!
Candidate A 30%
Candidate B 26%
"Whatever" 44%
Signed,
-Someone who took a basic statistics course.
JORGE CHAM © 2010
WWW.PHDCOMICS.COM

Probability Density Function (PDF)

- A PDF[2] is a *normalized*, distribution which tells you the probability of a finding your variable in some interval
- Random variable x[3] PDF $p(x)$ then

$$\text{Prob}(a < x < b) = \int_a^b p(x)dx \tag{1}$$

- The probability interpretation only makes sense if you must find your variable *somewhere*!

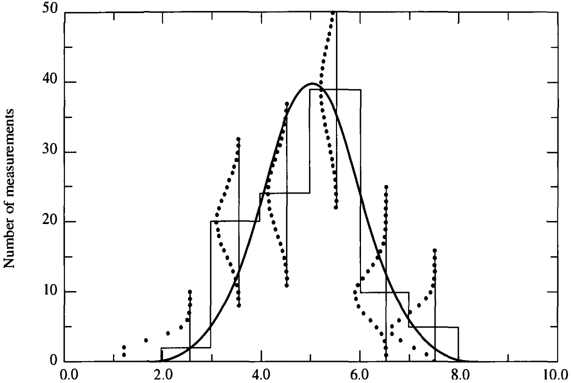
$$1 = \int_0^\infty p(x)dx \tag{2}$$

[2]Not Adobe's proprietary portable document format!

[3]Everything is a random variable! The energy deposited from a scattered photon (in Compton experiment), the number of entries in a specific bin of your MCA, the number of raindrops falling on your roof!

Histogram = Binned Data

- Data (number of counts) binned so that y-axis denotes number of counts in each bin. $h(x_i)$ is the number of counts in bin x_i



- “Integral” of data is total number of counts N
- Prediction shown as $Np(x_i)$
- Assumed Poisson distributions for each entry $h(x_i) \Rightarrow$ informs error bar

Error Bars

- Error bars are a tiny representation of the PDF for that data point
- Vertical ones usually denote $\pm 1\sigma$, which characterizes the underlying PDF for that particular data point
- Horizontal ones usually just indicate bin width (at least in Junior Lab)
- See Bevington Sections 1.2-1.3

χ^2 Test Statistic

- A χ^2 is a generally accepted variable[4] which can test the "goodness-of-fit," i.e. agreement between theory and experiment
- A χ^2 variable is a function of your data (reality), assumed error bars (uncertainty), and the PDF (theory)
- The definition is

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{h(x_i) - Np(x_i)}{\sigma_i(h)} \right)^2 \tag{3}$$

  If we know that each entry in a bin follows a Poisson process, then we can estimate $\sigma_i(h) = \sqrt{h(x_i)}$
- The expected value is

$$\langle \chi^2 \rangle = \nu = n - n_c = dof \tag{4}$$

  where n is the number of measurements (bins), n_c is the number of parameters
- See Bevington Section 4.3

Gradient Search

- Also known as method of *steepest descent*

$$\vec{\nabla}\chi^2 = \sum_{i=1}^{n} \frac{\partial \chi^2}{\partial a_i} \hat{a}_i \tag{5}$$

- "Go That Way"
- Pros: Good at getting close to minimum from far away fast
- Cons: Not great at finding minimum once it's in the neighborhood
- See Bevington Section 8.4

Levenberg-Marquardt

- Combines Grad Search with “Expansion Method”
- Expansion method finds an approximate analytic description of χ^2 near the minimum and uses this to find min
- Lev-Mar behaves like Grad Search far away from the minimum, then switches to be like the Expansion Method near the minimum
- Lev-Mar is tunable: λ controls the turning point
- See Bevington Sections 8.5-8.6

Tips & Tricks

- Make sure your errorbar vector `sig` does not contain zeros! (easy symptom: $\chi^2 = \infty$)
- Make all of your parameters the same order of magnitude
- Make sure there isn't some physical reason for your PDF not fitting your data, e.g. a constant background offset that your PDF doesn't take into account
- Try using Grad Search first, then use those output parameters as the starting parameters for Lev-Mar with a small λ parameter (more likely to find minimum)
- DON'T BE AFRAID TO LOOK THROUGH THE CODE!
- Remember what physics you're trying to extract: maybe fit to a smaller region-of-interest, e.g. you just want the mean of a Gaussian and don't care about the tail or background rate (subtract it off!)
- If all else fails, talk to me

References

1. Bevington, Bevington, Bevington
2. Bevington Section 4.3 for χ^2
3. Bevington Chapter 8 for methods! Sections 8.4 = Gradient Search, 8.5-8.6 = Levenberg-Marquardt
4. For a discussion of alternatives to $\sqrt{N}$ for Poisson errors: http://www-cdf.fnal.gov/physics/statistics/notes/pois_eb.txt

Example

Sum of Many Lorentzians (Mossbauer Experiment)
by S. Campbell (my partner) Junior Lab 2008