

Introduction to Data Fitting Ideas (Borrowed from 8.01T)

David Litster — Department of Physics

MIT

September 8, 2009

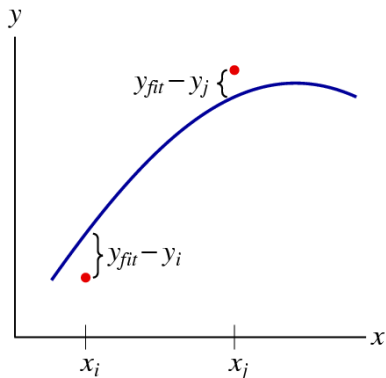
Why Fit Data?

Often when you have some measured data points $y_i(x_i)$ corresponding to independent variables x_i you would like to find a mathematical function $y_{\text{fit}}(x)$ that represents them. If there is a theoretical model that you think may explain your data then it will provide such a mathematical function.

- ▶ The model will have some parameters whose numerical values are unknown and can be adjusted.
- ▶ If the data are represented well by the model, then the values of the parameters in the model can be found from the fit.
- ▶ If the fit is good and the parameter values are physically reasonable, you might decide the model is correct.
- ▶ **How do you decide if the fit is good?**

Goodness of Fit

The goodness of fit is a measure of the distance between the data points and the function that is supposed to fit them.



$$TSE = \sum_{i=1}^N [y_{fit}(x_i) - y(x_i)]^2$$

where TSE is the total squared error between the data points and the equation that is supposed to fit them.

If the fit is to a straight line, $y(x_i) = A + Bx_i$, the parameters A and B are chosen to minimize a properly normalized TSE .

Goodness of Fit χ^2

Not all data points may have the same error (uncertainty) and one might expect more adjustable parameters would give a better fit. A quantity χ^2 is defined to take this into account. In normalized form (Bevington calls this χ^2_ν) this is

$$\chi^2_\nu = \sum_{i=1}^N \frac{[y_{\text{fit}}(x_i) - y(x_i)]^2}{(N - N_{\text{params}})\sigma_i^2}$$

where σ_i^2 is the variance (squared error) of the data point $y(x_i)$, N is the number of data points in the fit, and N_{params} is the number of adjustable parameters in the fit function.

Goodness of Fit χ^2 (cont'd)

- ▶ A good fit should have $\chi_\nu^2 \simeq 1$; if you find that χ_ν^2 is significantly less than 1, you have probably overestimated the uncertainties σ_j .
- ▶ Usually we will take all σ_i to be the same, and equal to a quantity σ .
- ▶ A fit is considered to be statistically better than another if it gives a χ_ν^2 that is two times smaller.
- ▶ The standard deviation of parameter A (σ_A), for example, is the change in parameter A from the optimum value (while optimizing all other parameters) that increases the *TSE* by about $N\langle\sigma_j\rangle$ or increases χ_ν^2 by about $1/N$.

Fitting Example: Human Vision

In 1942, S. Hecht, S. Shlaer, and M. Pirenne (HSP) published an article ¹ in which they investigated the detectability of light at the threshold of human vision. This classic paper in biophysics is discussed in detail in the excellent textbook *Physics With Illustrative Examples From Medicine and Biology: STATISTICAL PHYSICS* by George B. Benedek and Felix M. H. Villars [Springer-Verlag, 2000]. This discussion is taken from that book.

HSP found that the eye integrates for about 0.1 s and the response is determined by the number of photons that arrive in that time. They applied 1 ms flashes of light to a dark-adapted carrot-fed eye so that the light would fall on the most sensitive part of the retina, and measured the probability the light would be detected as a function of the number of photons incident on the cornea.

¹“Energy Quanta and Vision”, S. Hecht, S. Shlaer, and M. Pirenne, *Journal of General Physiology* **25**, 819 (1942)

Fitting Example: Human Vision (cont'd)

HSP measured the probability light would be seen as a function of n , the number of photons incident on the cornea. Their data are given in the table at the right.

n	$p(n)$
37	0.00
59	0.12
93	0.44
149	0.94
239	1.00

Their model had two assumptions; first that only $c n$ of the n photons will reach the retina to excite a conformational change in rhodopsin.

Second, it assumed that the number of rhodopsin molecules excited for a given number of photons obeyed Poisson statistics and that the light would be seen if at least m_0 molecules were excited.

Fitting Example: the Function

The model has two adjustable parameters: the attenuation c and the threshold for detection m_0 .

$$p(n) = \sum_{m=m_0}^{\infty} \frac{(cn)^m}{m!} e^{-cn} = 1 - \sum_{m=0}^{m_0-1} \frac{(cn)^m}{m!} e^{-cn}.$$

The fitting is slightly tricky, as m_0 must be an integer.

I wrote a program which found the c to minimize χ^2_ν for given values of m_0 . Fixing $\sigma = 0.02$ for each $p(n)$, the program gave the results at right.

m_0	c	χ^2_ν
7	0.0694 ± 0.0016	2.17
9	0.0901 ± 0.0014	1.13
12	0.1212 ± 0.0026	2.35

Fitting Example: Results

The model seems good, and it is interesting to learn that the human eye can detect as few as 10 photons.

