DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

www.nucleide.org

Curve fitting
general background

Philippe Cassette
Laboratoire national Henri Bequerel France

Laboratoire National LNHB Henri Becquerel

LNE Laboratoire national de métrologie et d'essais (LNE)

CURVE FITTING

- A set of *n* experimental points (with their uncertainties) is available
- We wish to adjust a mathematical function to these points

1/ How we do that?

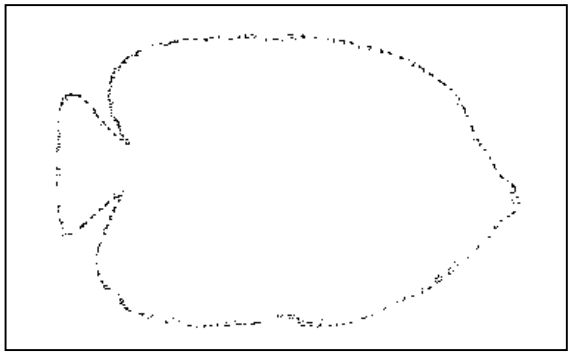
2/ What uncertainty can be attributed to the fitting function?

EXAMPLE OF CURVE FITTING

reality

experiment

sample



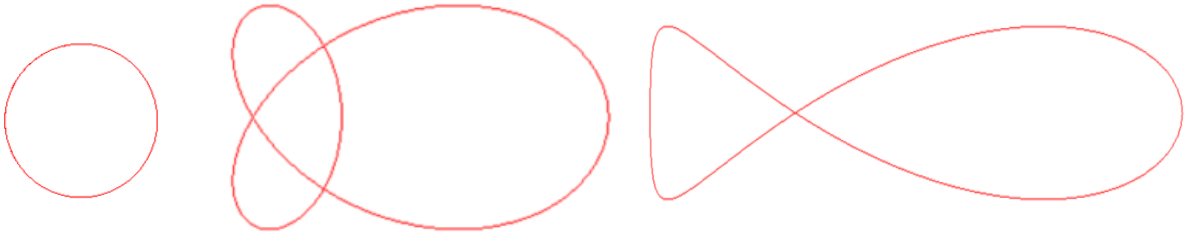
Fitting process:

- optimize *a* and *b* for the maximum likelihood of the model
- Calculate uncertainties

model

$$x = a\left(\cos\theta + b\cos\left(\frac{\theta}{2}\right)\right)$$

$$y = a\sin\theta$$



Various values of *a* and *b*

CURVE FITTING

Consider *n* experimental samples: $(x_i, y_i, s_{x_i}, s_{y_i}), \quad i = 1, n$

And a mathematical model with various parameters: $y = f(x, a, b, c...)$

- Find the best estimate of the parameters *a, b, c,…* for which the model is a good fit to the experimental points
- Evaluate the uncertainties of these parameters and of the fit
- Evaluate *a posteriori* the quality of the fit

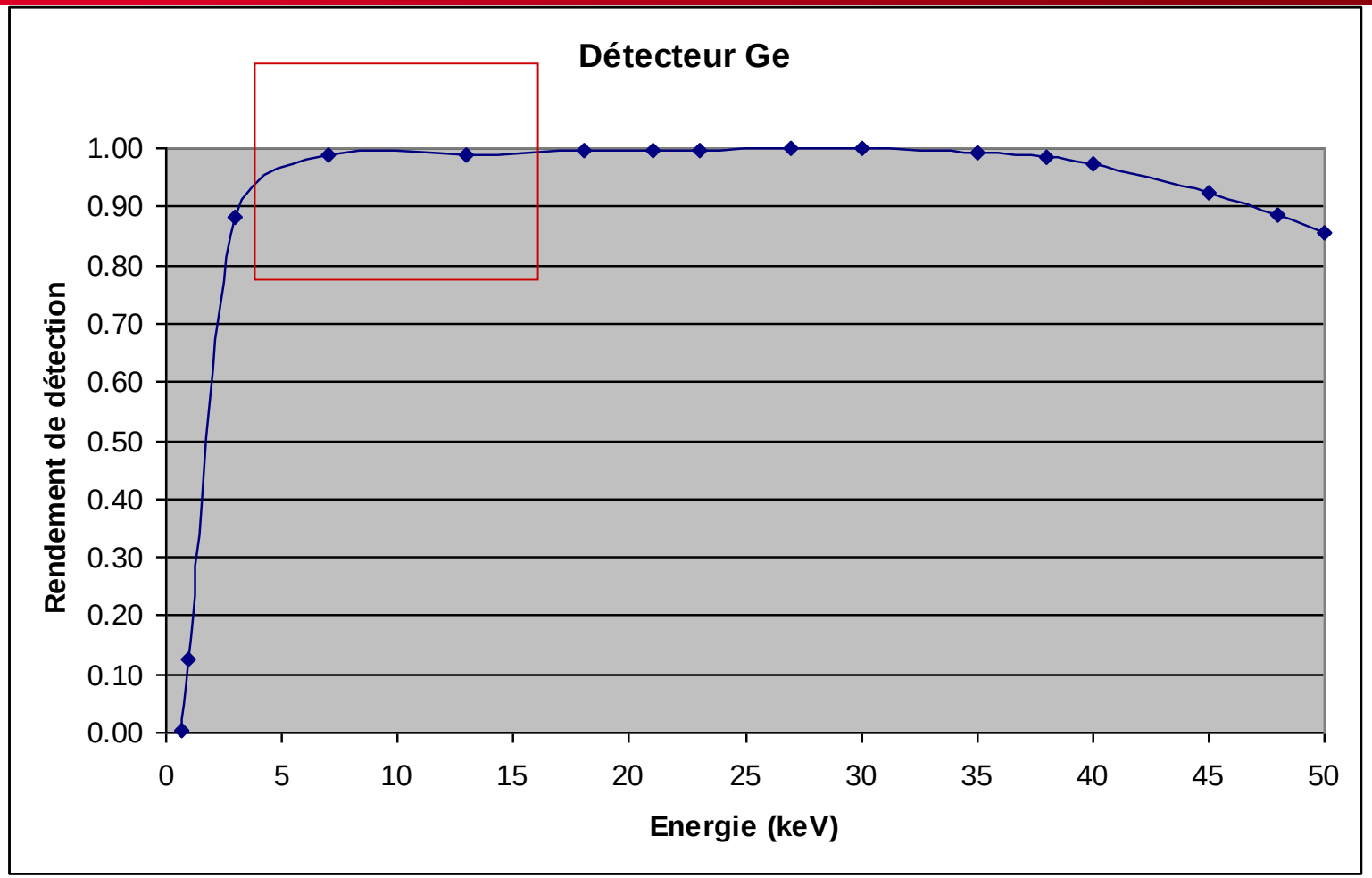
$$y_i(x_i) = f(x_i, a, b, c...) + \varepsilon_i$$

ε_i : random variable with (if possible) mean value ~ 0

GOAL OF A FIT

- To understand the physical laws governing an experiment
- To obtain a mathematical function easier to manage than a set of experimental points with individual uncertainties
- To interpolate between the points
- To extrapolate…

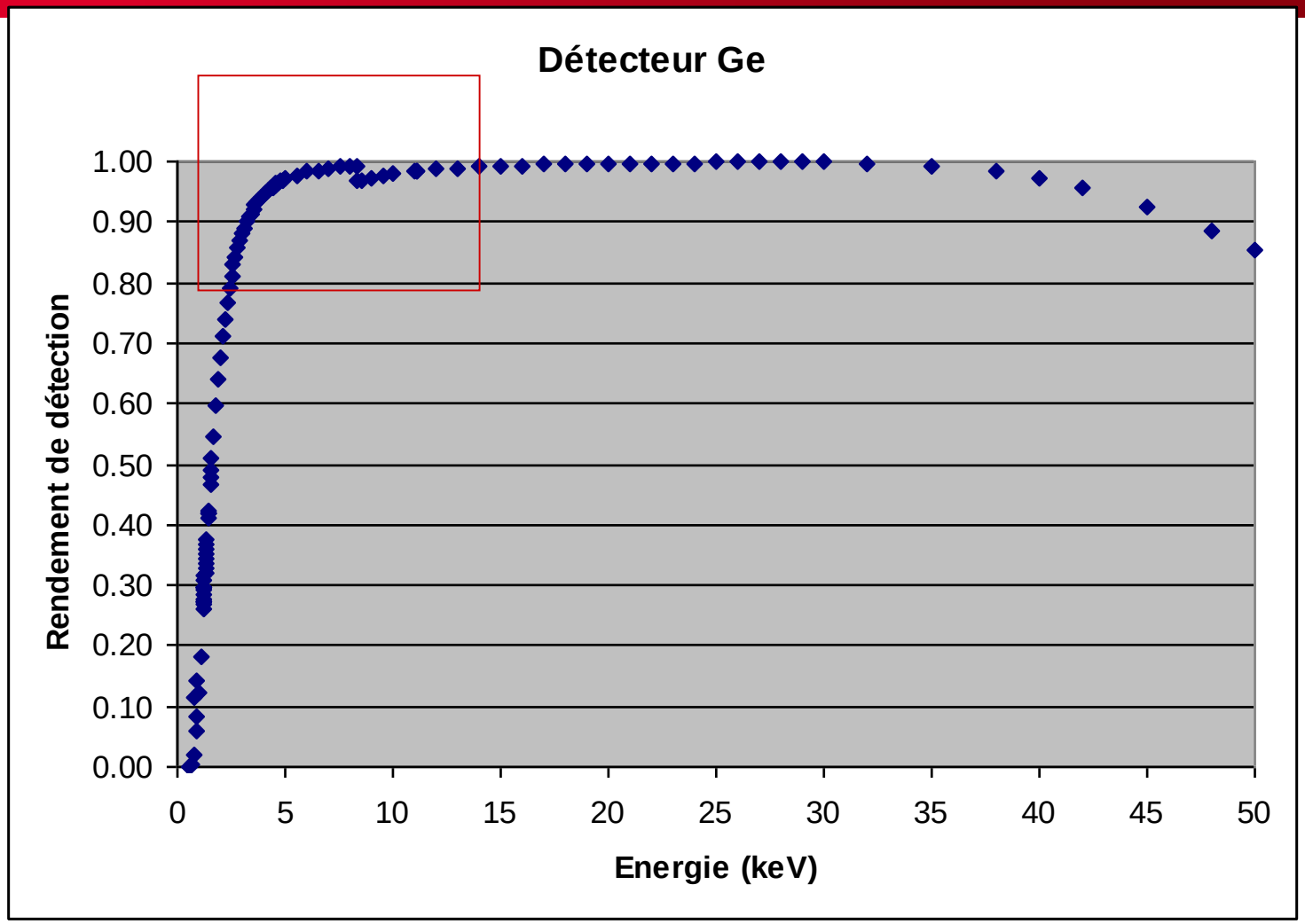
INTERPOLATION 1



8 keV, R= 0.99
13 keV, R= 0.99

→ Interpolation : 9 keV, R= 0.99

INTERPOLATION 2



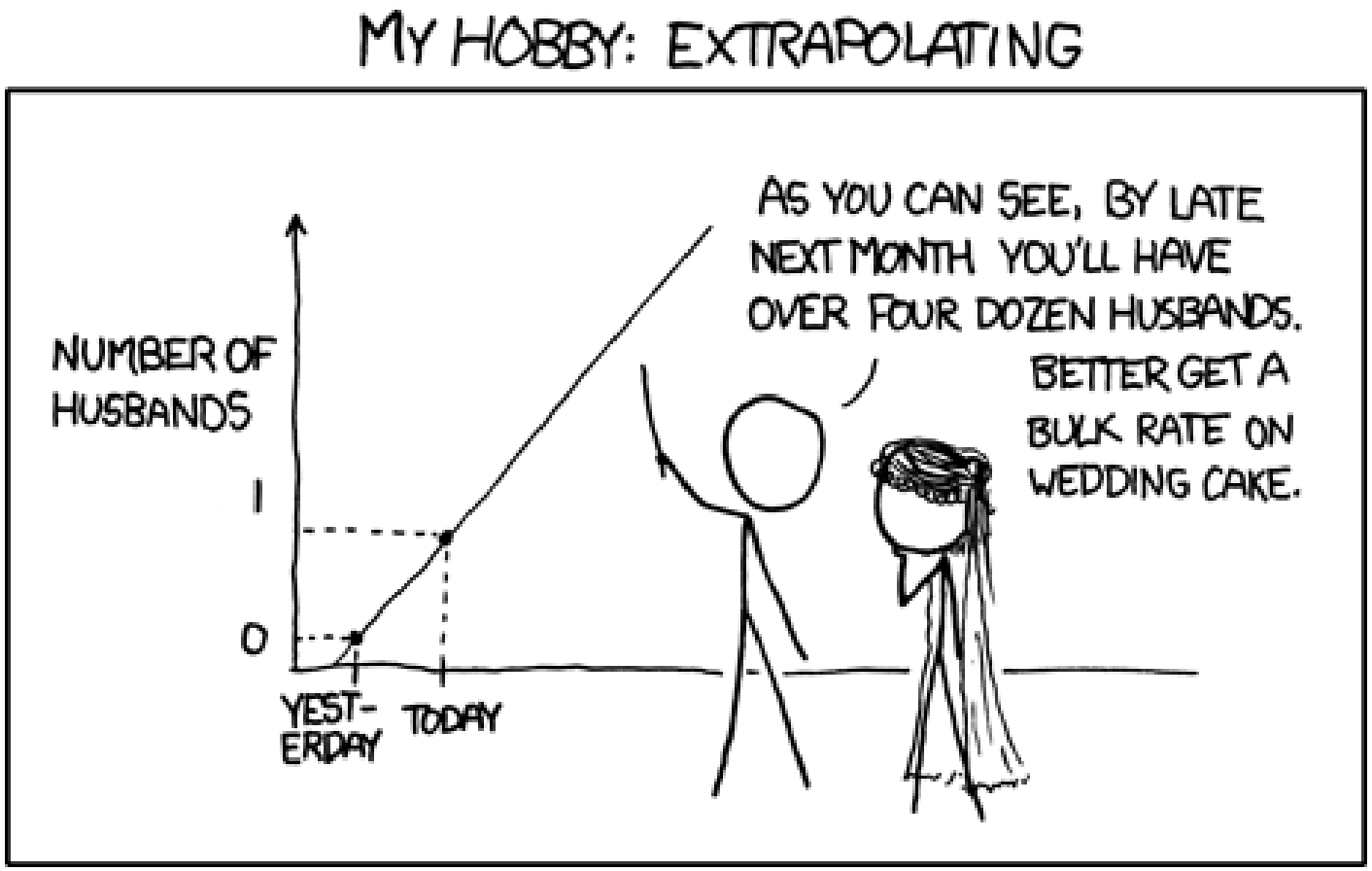
With a better sampling, it can be observed that R= 0.97 at 9 keV (Cu)

INTERPOLATION GENERAL RULES

Interpolation is applicable when there is no change of the physics between the experimental points, i.e.

- Smooth and continuous variation
- No discontinuity
- No phase transition

EXTRAPOLATION



BEST FIT, MAXIMUM LIKELIHOOD

Given a set of *n* experimental points (x_i, y_i) and a fitting function $y=f(x,a, b,\ldots)$ with parameters *a, b,…*

Calculate the probability, $P(y_i,f(x_i))$, that $f(x_i)$ is equal to y_i (within *dy)*

For this data set of n points, the likelihood, *L,* is the product of the probabilities that the value of each experimental point, y_i, is a statistical fluctuation of the function $f(x_i)$:

$$L = \prod_{i=1}^{n} P(y_i,\ f(x_i))$$

The best fit is obtained when L is maximized

GAUSSIAN FLUCTUATIONS

The probability that y_i is a Gaussian random fluctuation of $f(x_i)$ (within dy) is:

$$P \propto \exp\left[-\frac{1}{2}\left(\frac{y_i - f(x_i)}{u_i}\right)\right]^2 dy$$

The likelihood is:

$$L \propto \prod_{i=1}^{n}\left[\exp\left[-\frac{1}{2}\left(\frac{y_i - f(x_i)}{u_i}\right)^2\right] dy\right]$$

Maximizing L is equivalent to minimizing $-L$ or $-\log(L)$:

$$-\log L = \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{2u^2} - n\log(dy) + cst$$

n and dy being constant, the maximum likelihood corresponds to a minimum of:

$$\boxed{\sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{2u^2}}$$

Known as χ^2 minimization

IF ALL THE POINTS HAVE THE SAME UNCERTAINTY

Minimization of: $\sum_{i=1}^{n} (y_i - f(x_i))^2$

Least squares minimization

Thus: χ^2 or least squares fitting is only a maximum likelihood procedure if the experimental fluctuations are Gaussian

BUT IS REALITY NORMAL?

Gaussian distribution, mean M and standard deviation *u*

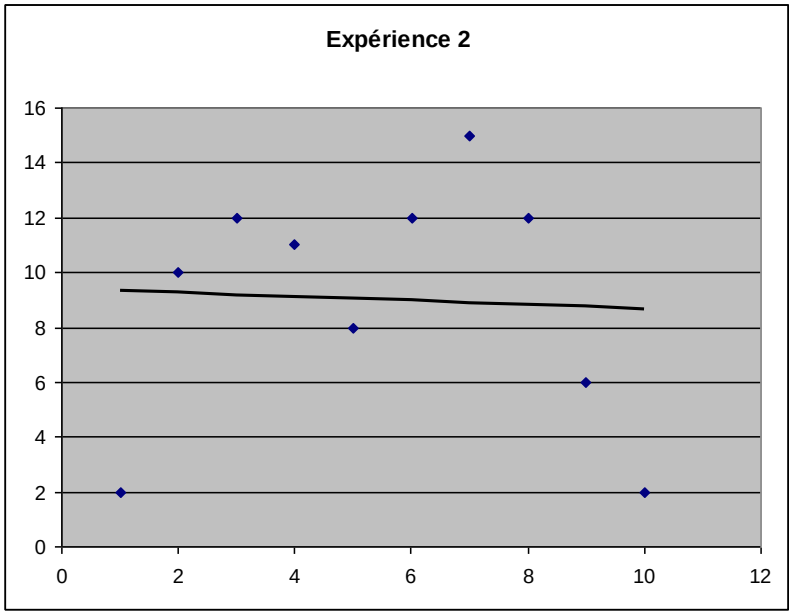
- 68 % of the population is in the ($M \pm 1u$) interval
- 95 % of the population is in the ($M \pm 2u$) interval
- 99,7 % of the population is in the ($M \pm 3u$) interval
- 99,9994 % of the population is in the ($M \pm 5u$) interval

But, the probability (frequency) to have an experimental point as far as 5*u* from the mean, even in an outstanding metrology lab, is likely to be larger than $3 \cdot 10^{-4}$ %....

Thus before using χ^2 or least squares fitting, better remove the outliers

WHAT ABOUT ADJUSTMENT QUALITY?

Example of linear regression



It is <u>always</u> possible to fit experimental points to any function… but sometimes it is better not to!

FIT OF A DETECTION EFFICIENCY CURVE IN GAMMA SPECTROMETRY

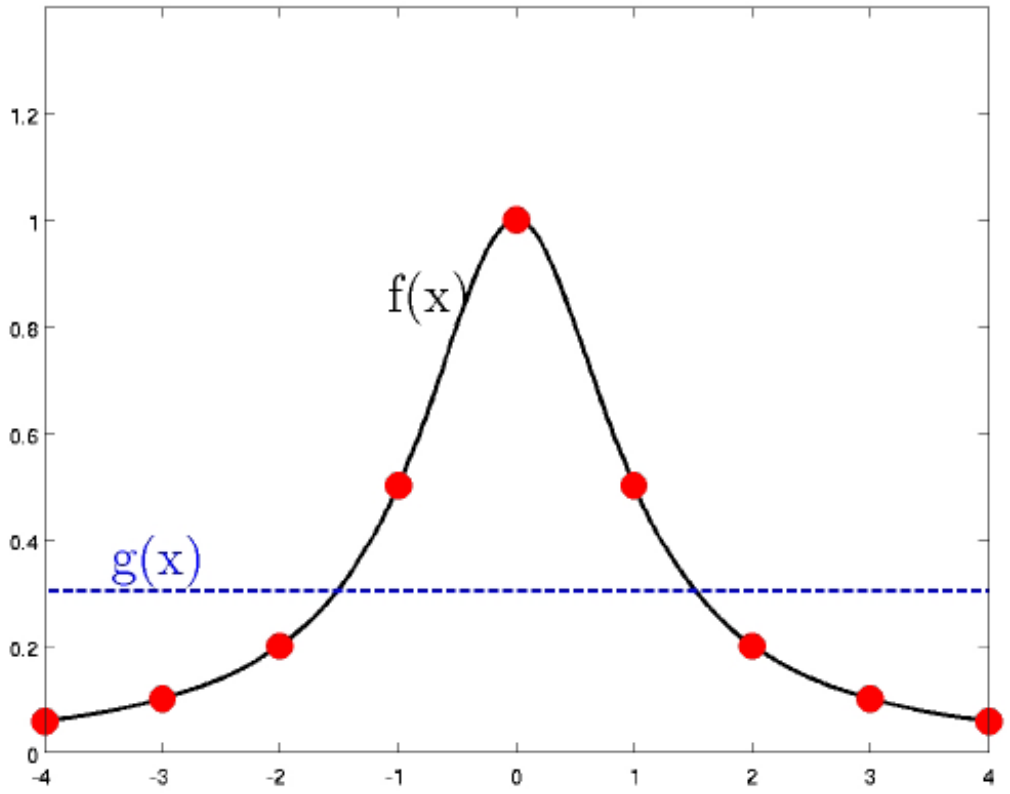
Physical model: efficiency decreases at low energy (window, dead-layer) and photons escape at high energy (with possible discontinuities due to K edges)
Physical model is seldom used

Arbitrary (albeit convenient) model: n^{th} order polynomial function

Known theorem: it is possible to exactly adjust an n^{th} order polynomial function to n points… but generally it is better to choose a lower order

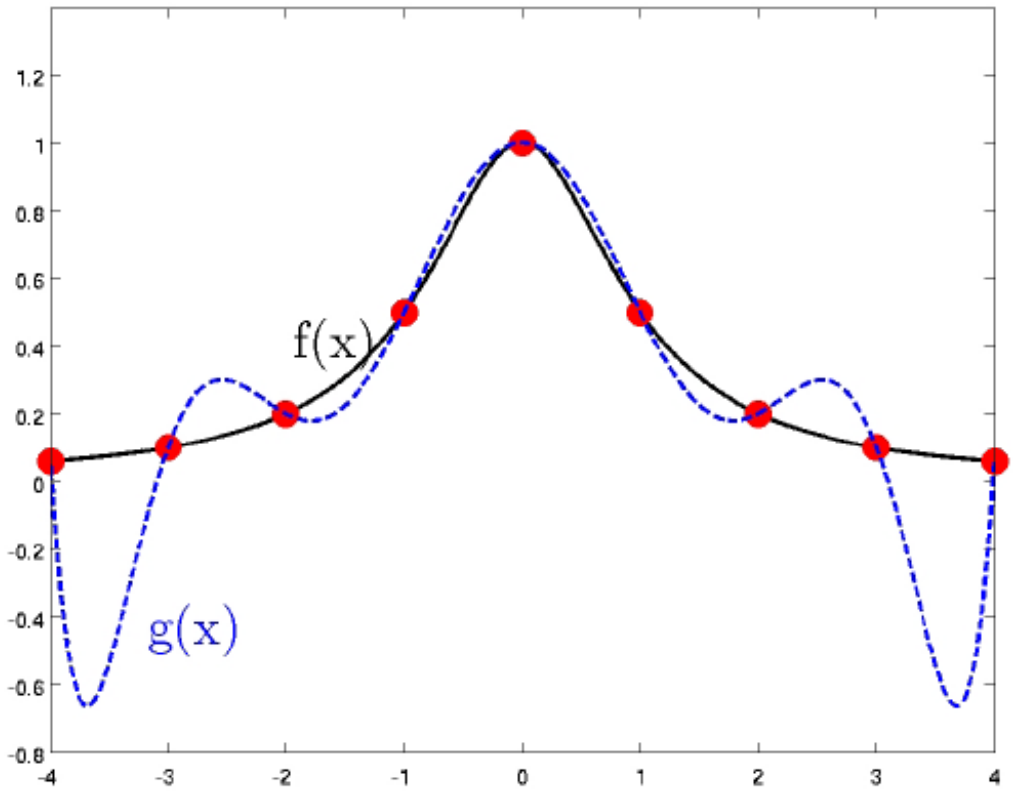
EXAMPLE OF TOO SMALL ORDER (1)

Poor adjustment!



EXAMPLE OF TOO LARGE ORDER (8)

Poor adjustment (but perfect residues)



HOW TO QUANTIFY THE GOODNESS OF FIT?

By calculating $Q(\chi^2, \nu)$ that the Chi-squared should exceed a particular value χ^2 by chance:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{u_{y_i}} \right)^2$$

The number of degrees of freedom, ν is the number of points minus the number of adjusted parameters

$$Q(\chi^2, \nu) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \int_{\chi^2}^{\infty} e^{-t} t^{\frac{\nu}{2}-1} dt$$

$Q > 0{,}1$: fair fitting
$0{,}1 > Q > 0{,}001$: acceptable fitting if uncertainties are underestimated
$Q < 0{,}001$: better try another fitting…

HOW TO QUANTIFY THE GOODNESS OF FIT?

• A too low value of *Q* reveals an inconsistency in the fitting process (model, or fitting criterion)

• *Q*=1 (too nice to be true), could indicate a large overestimation of the uncertainties, or synthetic results calculated from the model…

• Simple empirical rule: for a fairly good adjustment, the χ^2 value should be close to the degrees of freedom of the problem (i.e. if *n* points and a function with *k* parameters, $\nu = n - k$

More precisely, for large values of ν, the χ^2 distribution converges towards a Gaussian distribution of mean ν and standard deviation of $\sqrt{2\nu}$

GAMMA SPECTROMETER EFFICIENCY CURVE

Example where 4 calibration values are considered at 136.47 keV. What is the best strategy: consider the 4 values with associated uncertainties or calculate a mean value with associated uncertainty?

N°	Energie[keV]	Efficacité[%] (mesuré)	Efficacité[%] (Courbe)	Delta[%]	Sigma[%] (mesuré)	Sigma[%] (Courbe)
4	59.54	0.6125	0.5923	-3.3	2.02	0.35
5	59.54	0.5977	0.5923	-0.9	2.27	0.35
6	59.54	0.5909	0.5923	0.2	1.85	0.35
7	88.03	1.9924	1.9197	-3.6	2.40	0.61
8	88.03	1.9618	1.9197	-2.1	2.29	0.61
9	88.03	1.9535	1.9197	-1.7	2.46	0.61
10	122.06	2.6336	2.6372	0.1	1.22	0.76
11	122.06	2.6159	2.6372	0.8	0.55	0.76
12	122.06	2.6085	2.6372	1.1	1.28	0.76
13	122.06	2.5840	2.6372	2.1	1.15	0.76
14	136.47	2.7143	2.7117	-0.1	1.40	0.79
15	136.47	2.6445	2.7117	2.5	4.48	0.79
16	136.47	2.6193	2.7117	3.5	4.87	0.79
17	136.47	2.6155	2.7117	3.7	2.87	0.79
18	165.86	2.7529	2.6754	-2.8	1.28	0.80
19	165.86	2.7268	2.6754	-1.9	1.26	0.80
20	165.86	2.7145	2.6754	-1.4	1.19	0.80
21	165.86	2.6940	2.6754	-0.7	0.91	0.80
22	320.08	1.9195	1.9222	0.1	1.42	0.65
23	320.08	1.9174	1.9222	0.2	1.47	0.65
24	320.08	1.9098	1.9222	0.6	1.58	0.65
25	391.70	1.6819	1.6809	-0.1	1.02	0.59
26	391.70	1.6750	1.6809	0.3	1.02	0.59
27	391.70	1.6672	1.6809	0.8	1.01	0.59

If the 4 calibration points are independent (i.e. obtained using non-correlated sources) these 4 values can be taken into account, increasing the weight of this calibration point

The other possibility is to reduce these 4 values to 1, by calculating the mean and the standard deviation of the mean, but possible covariances must be taken into account in this calculation (some common uncertainty components could exist, if the sources are made in the same laboratory)

IF ALL CALIBRATION POINTS WERE TOTALLY UNCORRELATED

The same calibration curve is explored (and sampled) and there is a reduction of the global uncertainty, through the averaging process
In theory, the uncertainty vanishes when using an infinite (large) number of calibration points…

In practice, the calibration points are correlated (obvious example of multi-peaks gamma emitters), but the correlation could come from a common bias in primary calibration methods

It is thus prudent to consider a minimum value of the uncertainty (e.g. by considering the minimum uncertainty of a calibration point near the ROI)

LIMITATION OF THE AVERAGING PROCESS

The uncertainty resulting from the averaging of the results obtained using several sources cannot be lower than the intrinsic limitations of the uncertainty of a source (e.g. uncertainty due to the weighing process or intrinsic uncertainty of the primary measurement method used to calibrate the source)

In case of doubt, consider the use of an infinite number of sources. In this case, can the uncertainty converge towards zero, or could it be considered that a systematic (but unknown) bias can occur making unrealistic a zero uncertainty?

PRACTICAL CASE: RATIO OF GAMMA-RAY EMISSION INTENSITIES

Sometimes, the goal of gamma spectrometry is to give relative intensities of peaks (i.e. intensities compared to a reference peak supposedly known without uncertainty)

How to calculate relative uncertainties?

Origin of the problem: the reference emission is deduced from the analysis of the spectrum and thus is affected with an uncertainty. Is it necessary to consider it in the calculation?

ORIGIN OF THE UNCERTAINTY OF THE REFERENCE PEAK

Uncertainty:
- of the activity of the source
- of the detection efficiency
- from counting statistics, peak fitting and background subtraction

- the uncertainty of the source activity will impact all the emissions and thus is totally correlated with the uncertainty of the reference peak
- counting uncertainty and background uncertainty of the reference peak must be taken into account but, are likely to be negligibly small (if the reference is well chosen)
- uncertainty on the detection efficiency must be taken into account but the correlation must be taken into account (especially if the peaks are close to the reference peak)
- Peak fitting uncertainty?

THUS

The reference peak is normalized, but, as it is the result of a measurement, it is only known by its estimator and has an intrinsic uncertainty

Thus this uncertainty must be considered in the calculation of the intensity ratio using the formula of propagation of variances (or a Monte Carlo method)

Common uncertainty components for all peaks must not be taken into account (e.g. activity of the source)

For peak fitting and efficiency curve uncertainties, there is no general rule

CURVE FITTING OF POINTS WITH UNCERTAINTIES IN X AND Y

The Monte Carlo method is probably the easiest way

- Consider n points (x_i,u_{xi},y_i,u_{yi}), $i=1,n$
- The probability density function is assumed (Gaussian, uniform,…)

Calculation of a set of random fluctuations $((x_1,y_1), (x_2,y_2)…(x_n,y_n))$
Curve fitting on each set $f(x,a,b,c…)$ giving a, b, c, …

The calculation is repeated k times and we get k values of a, b, c,…

Calculation of the arithmetic mean of the k values of each parameter

Calculation of the variance/covariance matrix

→ standard uncertainty of each parameter
covariances between the parameters

UNCERTAINTY OF THE FIT

Example efficiency curve, Gaussian fluctuations of the points:

- Calculate a centered Gaussian G(0,1)
- Calculate the fluctuations of the experimental points

  $eff_1 = eff_0 + s_{eff0} * G(0,1)$
- Calculate the polynomial, e.g. $eff_1 = a + b*E + c*E^2 + d*E^3$
- Repeat this calculation n times
- Calculate the mean, the standard deviation of the mean, the covariances

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$s_a = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2}$$

idem for b,c…

$$s_{a,b} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})$$

Laboratoire National
LNHB
Henri Becquerel

THANK YOU