Fitting data: an introduction

Frank Verbunt

MSSL meeting on high-resolution X-ray spectroscopy

17 March 2009

Outline

- Parent distributions: concept and examples
- Central limit theorem, Gaussian errors and χ^2
- Methods for Gaussian errors: linear, non-linear
- General methods: amoebe, genetic algorithms
- Binning

Excerpted from full notes:

`www.astro.uu.nl/~verbunt/onderwijs/observe/lnotes.pdf`

based a.o. on Bevington

Parent distribution: concept

How not to...

- what is the probability that during this lecture we are hit by a meteorite?
- there are two possibilities: yes/no
- thus the probability is 50%

How to...

Determine

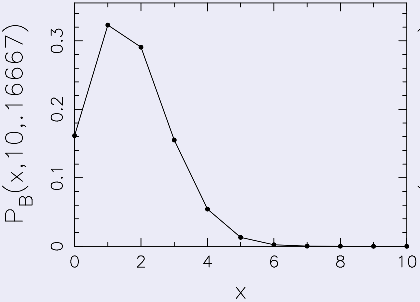
- possible outcomes
- their (relative) probabilities

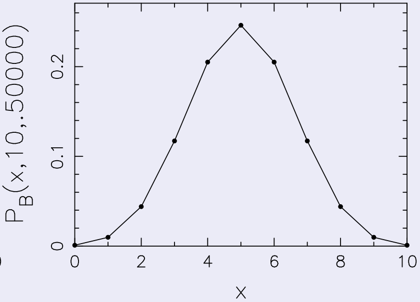
The combination is the parent distribution. It is never know exactly, always only approximately

Parent distribution: binomial

- expected: μ photons in time T, divide T in n slots
- each slot has probability $p = \mu/n$ to receive photon
- with n trials the probability of k hits and thus $n - k$ empty is

$$P_B(k, n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

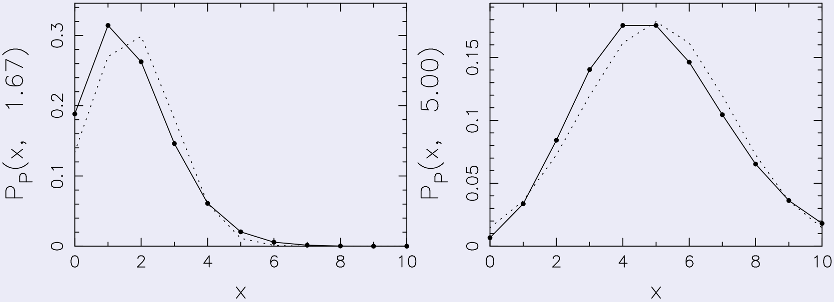




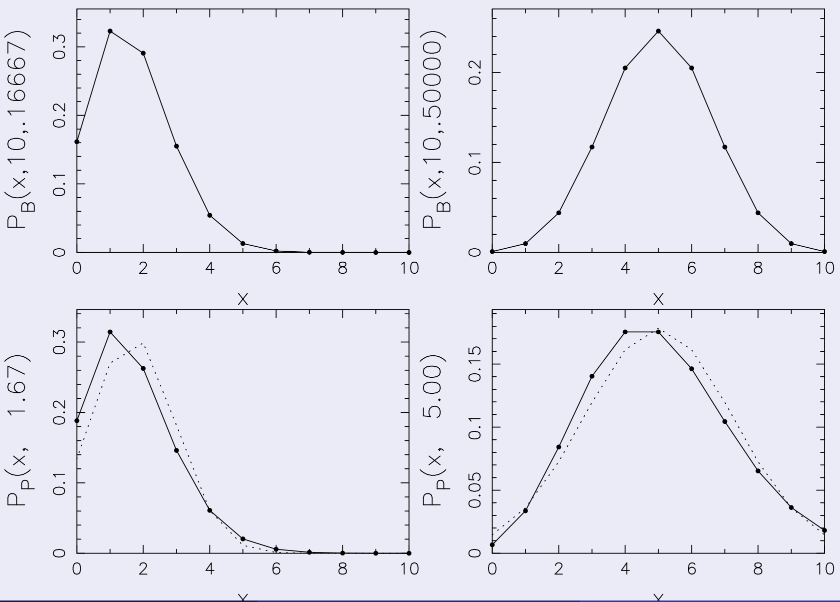
Parent distribution: Poisson

- expected: μ photons in time T, divide T in n slots
- each slot has probability $p = \mu/n$ to receive photon
- to avoid 2 photons in 1 trial, take limit $n \to \infty$ with np constant

$$P_P(k,\mu) = \frac{\mu^k}{k!} e^{-\mu}$$



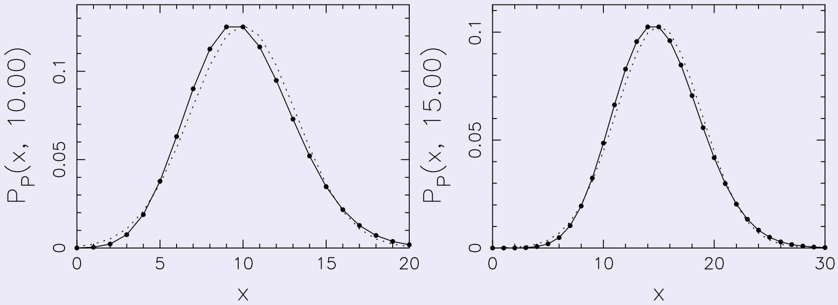
Parent distribution: Binomial to Poisson



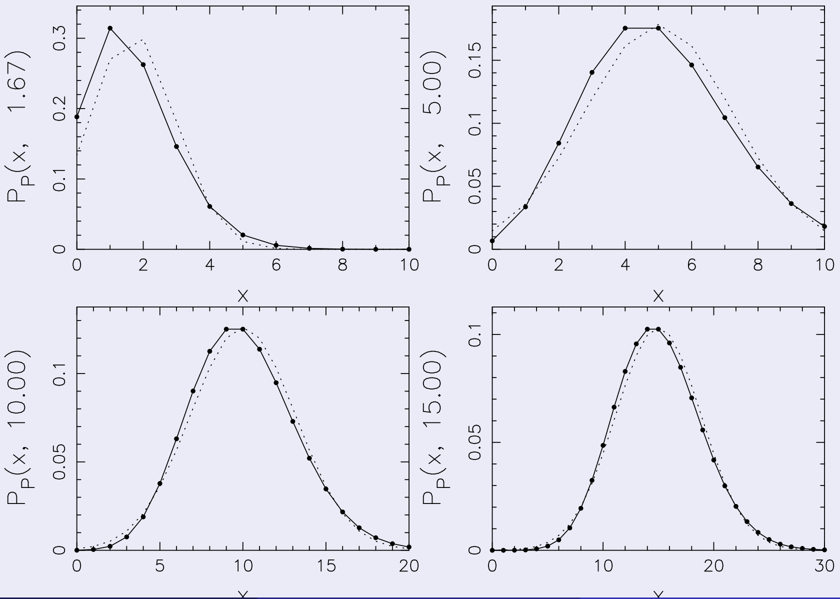
Parent distribution: Gauss

- expected value μ photons in time T
- for large μ the Poisson distribution is well approximated with the Gauss distribution

$$P_G(x,\mu,\mu) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\mu^2}$$



Parent distribution: Poisson to Gauss



Parent distribution: when to use which one

Gauss and normal

$$G(x,\mu,\sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

coordinate transformation: $z = (x-\mu)/\sigma$ gives normal distribution:

$$P_G(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right]$$

when to use

- binomial: each trial has outcome yes or no
- Poisson: each trial has range of possible outcomes
- Gauss: replaces Poisson for large expectation value
- for photon counts Gauss is never exact: in particular large deviations are more likely in Poisson

Central limit theorem

Concatenation of uncertainties

- The central limit theorem states that a sequence of various distributions applied consecutively will approximate a Gaussian
- For this reason and for its computational simplicity, the assumption of Gaussian error distributions is often used

How do we know?

- once we have a fit, we can plot distribution of the errors and check whether it looks Gaussian
- in general the errors are NOT Gaussian
- but the fit obtained by assuming they are is often not far wrong...
- how far is too far?

Gaussian errors and χ^2 minimalization (Press et al.)

- measurements y_i with associated Gaussian errors σ_i
- i.e. each drawn from a Gaussian around model value y_m
- probability for one measurement y_i, in an interval Δy, is

$$P_i \Delta y = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-(y_i - y_m)^2}{2\sigma_i{}^2}} \Delta y$$

The overall probability of a series is:

$$P(\Delta y)^N \equiv \prod_{i=1}^{N} (P_i \Delta y) = \frac{1}{(2\pi)^{N/2} \prod_i \sigma_i} \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - y_m)^2}{\sigma_i{}^2}\right] \Delta y^N$$

The highest probability P is that for which

$$\chi^2 \equiv \sum_{i=1}^{N} \chi_i{}^2 \equiv \sum_{i=1}^{N} \frac{(y_i - y_m)^2}{\sigma_i{}^2}$$

Gaussian errors and χ^2 minimalization

the observed χ^2

- N measurements y_i at measurement points x_i
- each y_i is drawn from a Gaussian
- i.e. each $\chi_i \equiv (y_i - y_m)/\sigma_i$ is a draw from the normal distribution
- square all χ_i's and add: $\chi^2 \equiv \sum_{i=1}^{N} \chi_i^2$

In a fit with N measurements and M fit parameters we have $\nu \equiv N - M$ independent draws

χ^2-distribution

- Simulate a measurement by randomly choosing a set of ν values y_i at x_i
- this is called a realization
- compute for many realizations the χ^2, to obtain the χ^2-distribution for ν
- for a Gaussian, this can be done semianalytically
- $\nu \equiv N - M$ is called 'degrees of freedom' or d.o.f.

Gaussian errors and χ^2 minimalization

Semi-analytic

- consider the incomplete Gamma function:

$$Q(a,x) \equiv \frac{1}{\Gamma(a)} \int_x^\infty t^{a-1} e^{-t} dt$$

- the fraction of $\chi^2 > \chi_o^2$ is given by Q with $a = 0.5(N-M)$ and $x = 0.5\chi_o^2$
- the probability of obtaining a χ^2 as observed or bigger is given hereby

Rule of thumb

- if $\nu \equiv N - M$ is large, then we expect roughly
- $\chi^2 \simeq N - M; \chi_r^2 \simeq 1$
- with a spread $\sqrt{2(N-M)}$

if χ^2 high, Q very small

- the model is wrong
- σ_i under-estimated
- errors not Gaussian

or a combination of these...
hence: tolerance of 'low' Q, e.g. 0.05 or 0.01

Gaussian errors and χ^2 minimalization

Effect of non-reporting

- a person has guessed a 6 digit number correctly
- the probability is 1 in 10^6
- so that person is special!
- unless she/he is one of a million persons who guessed...

If only significant results are published, the significance of published results will be over-estimated

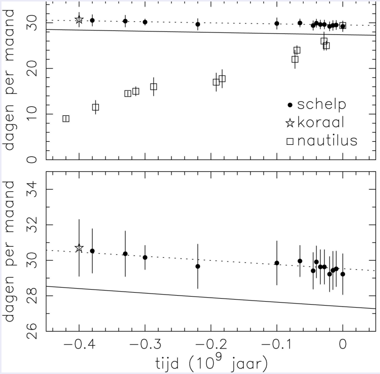
A good fit

consists of three parts

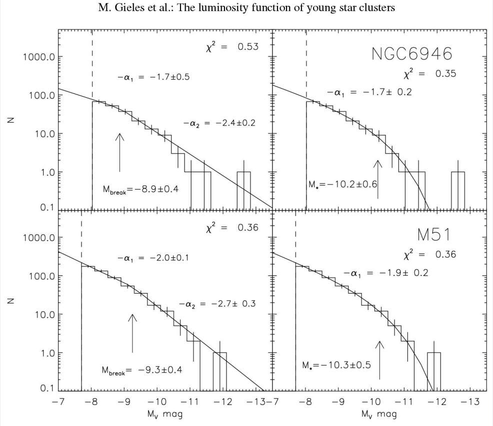
- the best value parameters
- the uncertainties on these parameters
- the probability that the model describes the data (either χ^2 and d.o.f. or Q)

See what is wrong, without knowing details...

Number of days/yr



Number of systems vs. M_V



χ^2 minimalization with linear dependence on model parameter: example weighted average

model $y_m = a$. Minimize χ^2 with respect to a:

$$\frac{\partial}{\partial a}\left[\sum_{i=1}^{N}\frac{(y_i - a)^2}{\sigma_i{}^2}\right] = 0 \Rightarrow \sum_{i=1}^{N}\frac{y_i - a}{\sigma_i{}^2} = 0 \Rightarrow a = \frac{\sum_{i=1}^{N}\frac{y_i}{\sigma_i{}^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i{}^2}}$$

a is a function of the variables $y_1, y_2, \ldots$. *If the measurements y_i are not correlated*, we find the variance for a from

$$\sigma_a{}^2 = \sum_{i=1}^{N}\left[\sigma_i{}^2\left(\frac{\partial a}{\partial y_i}\right)^2\right] = \sum_{i=1}^{N}\left[\sigma_i{}^2\left(\frac{1/\sigma_i{}^2}{\sum_{k=1}^{N}(1/\sigma_k{}^2)}\right)^2\right] = \frac{1}{\sum_{i=1}^{N}(1/\sigma_i{}^2)}$$

In general: if y_m is a linear function of model parameters a_k $(k = 1, M)$ the summations can be done without knowing a_k, and the solution is found directly

Gaussian errors and χ^2 minimalization

linear: straight line

$$y_m(x_i, a, b) = a + bx_i$$

minimize χ^2:

$$\frac{\partial \sum_{i=1}^{N} [(y_i - a - bx_i)/\sigma_i]^2}{\partial a} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{{\sigma_i}^2} \right) = 0 \Rightarrow$$

$$\sum_{i=1}^{N} \frac{y_i}{{\sigma_i}^2} - a \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} - b \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} = 0$$

again: sums can be done without knowing a, b: direct solution

Nonlinear example: sine

$y_m = \sin(ax)$ Minimize χ^2:

$$\frac{\partial \chi^2}{\partial a} = 0 =$$

$$-2 \sum_{i=1}^{N} \frac{[y_i - \sin(ax_i)]x_i \cos(ax_i)}{{\sigma_i}^2}$$

One cannot do the sums without a value for a. Hence the solution must be found iteratively

χ^2 minimalization with Levenberg-Marquardt

one dimension

far from minimum use

$$a_{n+1} = a_n - K\frac{\partial\chi^2}{\partial a}$$

Close to minimum approximate

$$\chi^2(a) = p + q(a - a_{min})^2$$

$$\partial\chi^2/\partial a = 2q(a - a_{min})$$

$$\partial^2\chi^2/\partial a^2 = 2q$$

$$\Rightarrow a - a_{min} = \frac{\partial\chi^2/\partial a}{\partial^2\chi^2/\partial a^2}$$

more dimensions $y_m(x, \vec{a})$

$$\chi^2(\vec{a}) \simeq p - \vec{q}\cdot\vec{a} + \frac{1}{2}\vec{a}\cdot\vec{\vec{D}}\cdot\vec{a}$$

$$\frac{\partial\chi^2}{\partial a_k} = -2\sum_{i=1}^{N}\frac{[y_i - y_m]}{\sigma_i{}^2}\frac{\partial y_m}{\partial a_k} \equiv -2\beta_k$$

$$\frac{1}{2}\frac{\partial\chi^2}{\partial a_k\partial a_l} \equiv \alpha_{kl} =$$

$$\sum_{i=1}^{N}\frac{1}{\sigma_i{}^2}\left[\frac{\partial y_m}{\partial a_k}\frac{\partial y_m}{\partial a_l} - [y_i - y_m]\frac{\partial^2 y_m}{\partial a_k\partial a_l}\right]$$

thus $\beta_k = \lambda\alpha_{kk}\delta a_k$ or $\beta_k = \sum_{l=1}^{M}\alpha_{kl}\delta a_l$

χ^2 minimalization with Levenberg-Marquardt

matrix equation

$$\beta_k = \sum_{l=1}^{M} \alpha_{kl} \delta a_l$$

with

$$\alpha_{kl} = \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} \left[\frac{\partial y_m}{\partial a_k} \frac{\partial y_m}{\partial a_l} \right]$$

iterate computation of δa_i until minimum of χ^2 is reached. If a_i not correlated, then

$$\delta\chi^2 = \delta\vec{a} \cdot \vec{\vec{\alpha}} \cdot \delta\vec{a} = \alpha_{kk} \delta {a_k}^2$$

Problems

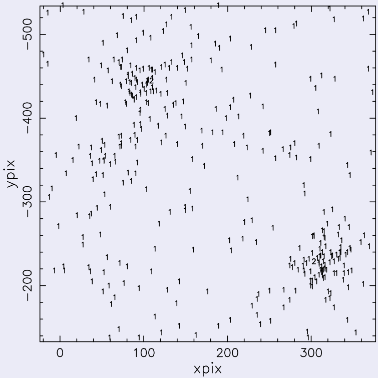
- requires reasonably close first estimate
- may converge to local minimum: try different starting solutions
- when number of parameters big: matrix very large

First derivative

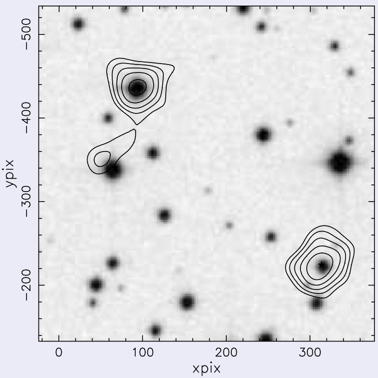
- when not analytic
- then compute numerically (with small step in a_i)

Example: a rosat image frame, with 0, 1, 2

counts per pixel



smoothed + optical



clearly, Gaussian statistics don't apply. what to do?

Poisson errors and maximum-likelihood

n_i photons when m_i expected

$$P_i = \frac{{m_i}^{n_i} e^{-m_i}}{n_i!}$$

Maximize overall probability $L' \equiv \prod_i P_i$:

$$\ln L' \equiv \sum_i \ln P_i =$$

$$\sum_i n_i \ln m_i - \sum_i m_i - \sum_i \ln n_i!$$

or equivalently minimize

$$\ln L \equiv -2\left(\sum_i n_i \ln m_i - \sum_i m_i\right)$$

Comparing models

- models A and B
- number of fitted parameters n_A, n_B
- likelihoods $\ln L_A$, $\ln L_B$

$$\Delta L \equiv \ln L_A - \ln L_B$$

is χ^2 distribution with $n_A - n_B$ d.o.f. (for a sufficient number of photons)

- probability of best solution from simulations

General fitting methods: amoebe

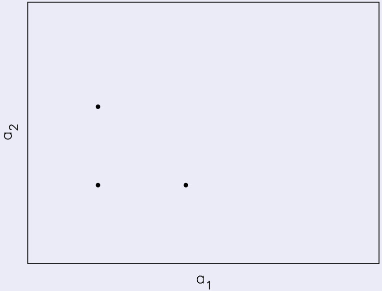
When?

- when number of parameters of χ^2 too big
- when errors not Gaussian

General

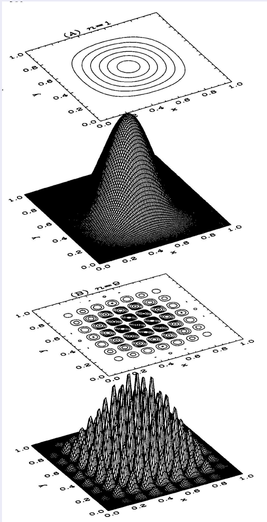
- do not use derivative: easier to programme, esp. for complicated derivative
- no fast convergence
- errors must be computed explicitly by changing parameter of best solutions

Amoebe in 2-d



- find worst point and move it
- repeat (also with other points) until minimum reached

General fitting methods: genetic algorithm



χ^2 or L varies erratically

$$f(x,y) =$$
$$[16x(1-x)y(1-y)\sin(n\pi x)\sin(n\pi y)]^2$$

- varies smoothly for $n = 1$ (top)
- varies wildly for $n = 9$ (bottom)
- Levenberg-Marquardt fails miserably...
- surprisingly, `amoebe` works well

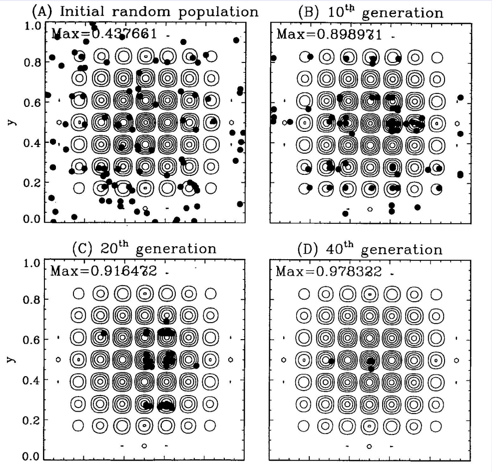
Charbonneau

General fitting methods: genetic algorithm

Encoding:	Ph(P1)	x=0.14429628	y=0.72317247	[01]
	Ph(P2)	x=0.71281369	y=0.83459991	[02]
		↓ 71281369	↓ 83459991	[03]
	Gn(P2)	7128136983459991		[04]
Breeding:	Gn(P1)	1442962872317247		[05]
	Gn(P2)	7128136983459991		[06]
(a) Crossover (gene=4):				
		144 | 2962872317247		[07]
		↓↑		
		712 | 8136983459991		[08]
		144 | 8136983459991		[09]
		712 | 2962872317247		[10]
	Gn(01)	1448136983459991		[11]
	Gn(02)	7122962872317247		[12]
(b) Mutation (Offspring=02, gene=10):				
	Gn(02)	7122962872317247		[13]
		712296287 [2] 317247		[14]
		712296287 [8] 317247		[15]
	Gn(02)	7122962878317247		[16]
Decoding:	Gn(02)	7122962878317247		[17]
		71229628 ↓	78317247 ↓	[18]
	Ph(02)	x=0.71229628	y=0.78317247	[19]
	Ph(01)	x=0.14481369	y=0.83459991	[20]

- two parameters: x,y
- paste digits together to make ‘animal’
- make generation of e.g. 100 animals
- compute goodness of fit χ^2 or L for each animal
- assign breeding probability according to goodness of fit
- breed with changeover and mutation

General fitting methods: genetic algorithm



Properties

- fitness (i.e. breeding probability) on ranking (e.g. rank n has probability $\propto 1/n$)
- elitism: keep best solution(s)
- mutation rate not too high, esp. in beginning
- final convergence slow
- fun variant: bad sheep

Some remarks on binning

One should not bin too much
Rue-of-thumb: 3 bins per FWHM resolution of instrument

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{N_i - M_i}{\sigma_i} \right)^2$$

with $\sigma_i = \sqrt{N_i}$. Split each bin in p bins: $N'_i = N_i/p$, $M'_i = M_i/p$, $\sigma'_i = \sqrt{N_i/p} = \sigma_i/\sqrt{p}$ hence

$$\chi'^2 = \chi^2/p$$

with smaller χ^2 and larger N, the quality of fit Q will be bigger. $\Rightarrow$ by oversampling an unacceptable fit may be made acceptable

Gaussian

- the Fourier transform is also Gaussian
- small bins are high spatial frequencies
- but with small number of photons we have no info on high spatial variability
- $\Rightarrow$ FT components at high frequencies are spurious (noise)

Frank Verbunt

MSSL meeting on high-resolution X-ray spectroscopy

17 March 2009

Conclusion

Statistics is not all that difficult
Combine some basic knowledge with common sense

More explanation and references in Lecture Notes:
www.astro.uu.nl/˜verbunt/onderwijs/observe/lnotes.pdf