Bayesian methods for galaxy evolution.

i.e. chap 12 of

Bayesian Methods for Cosmology

Cambridge University Press

Stefano Andreon
INAF-Oss. Di Brera
stefano.andreon@brera.inaf.it

Plan of the talk

- Why we need stats
- When we mostly need stats
- Why what-it-is-best-known doesn't work when we mostly need it (and work well instead when we don't need it)?
- What are numbers returned by the most known statistical tests?
- If prior are bad, why astronomers talk about Malmquist or Eddington bias?
- A simple example of a non-academic case: determination of the variance (width) of a distribution in presence of a background population
- Mixture of regressions, i.e. colour-magniture relation, Fundamental Plane, cluster scale relationship, Tully-Fisher relation, Magorrian relation, etc.

This talk is an open discussion.

Don't esitate: ask me when I am unclear, when you want to make comments or to pose questions, to suggest a solution, or … when you want

I'm provocative, but serius.

Caveat: talk for astronomers.

Stages in Astronomy Experiments

Stage	How	Example	Considerations
OBSERVE	Carefully	Experiment design, exposure time (S)	What? Number of objects, Type? (S)
REDUCE	Algorithms	calibration files QE,RMF,ARF,PSF (S)	data quality Signal-to-Noise (S)
ANALYSE	Parameter Estimation, Hypothesis testing (S)	Intensity, positions (S)	Frequentist Bayesian? (S)
CONCLUDE	Hypothesis testing (S)	Distribution tests, Correlations (S)	Belivable, Repeatable, Understandable? (S)
REFLECT	Carefully	Mission achieved? A better way? We need more data! (S)	The next Observations (S)

Wall & Jenkins (2003)

In which occasions we really need stats?

In most papers the measured effect is not overwhelming (obvious results have been already published before!). Most of the times we are working with small samples, tiny effects or complex problems and a number of subtle issues. It is precisely in these cases that we are interested in statistics.

We cannot avoid to work in this 'gray area':

1) we need to convince TAC about our good proposal, we just have an hint of something interesting, and we need better data
2) we cannot ask to a transient event (e.g. GRB) to become permanent enough to allow us to move our measurement in the 'statistical significant' area
3) sometime a larger sample does not exist, having already observed the whole universe (in the microwave, for example) or all galaxies in a given group
4) sometime the 'problem' has nothing to do with sample size, and thus do not disappear enlarging the sample (e.g. 60% of the hundred thousand galaxies of the nearby universe have negative SFR)
5) if we have just two determination of an important physical quantity (SFR at z=6) why we cannot combine them to take profit of both?
6) if I don't publish my favorite result someone else will publish it (and will get credit for the result)!

It is in this gray area that many times standard methods fail.

How to be wrong: I, defs

Mixture distributions arise when data come from two distribution:

- signal intensity or surface brightness and background
- red and blue galaxies
- two objects in the field of view of my instrument
- galaxy SED (two stellar populations)
- Galaxies in the space of their parameters (r_e, μ_e, concentration, σ_v, etc)

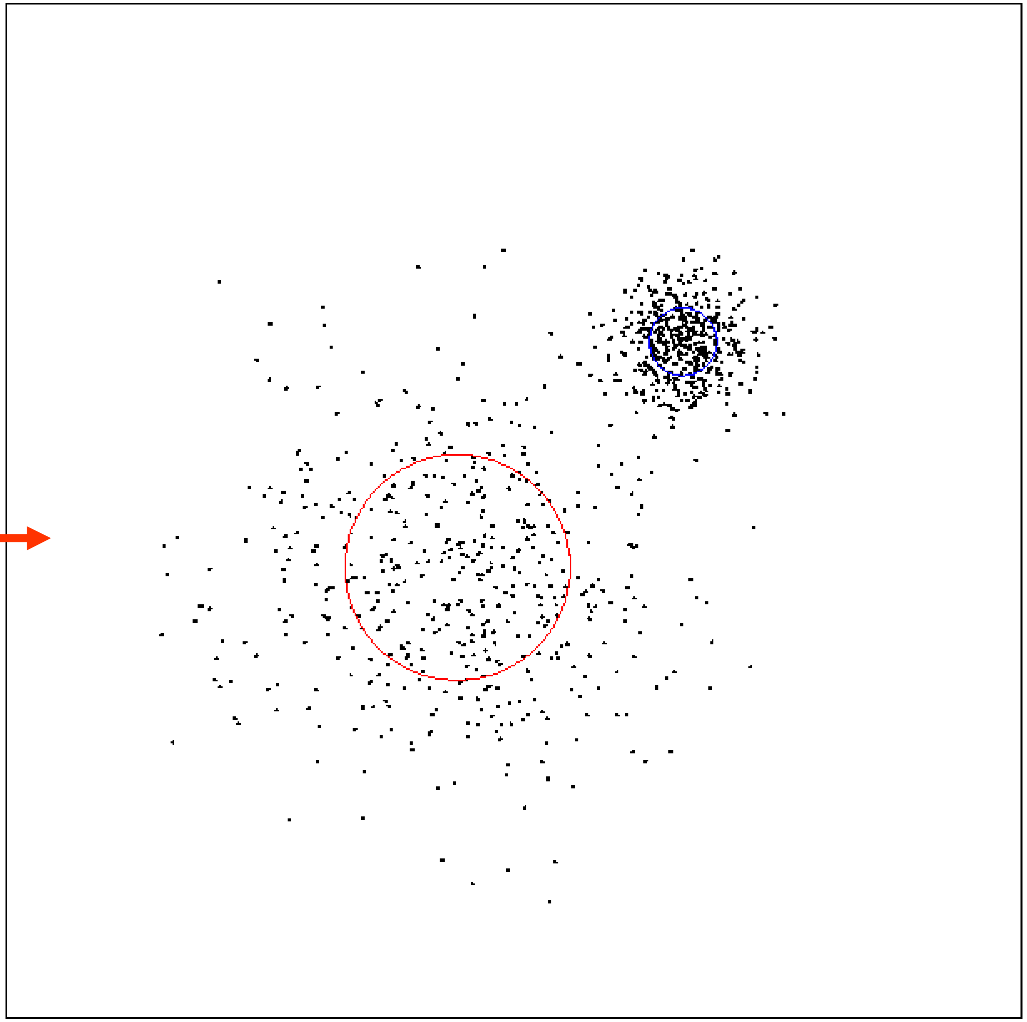
Mixture are ubiquitous in astronomy

How to be wrong: II, a toy example

The problem:

Determine parameters of the sum (mixture) of two 1-D gaussians of unknown (to be determined) center and width.

a 2 D example:



Thesis: the most likely value of the parameters are wrong.

How to be wrong: III, the maths

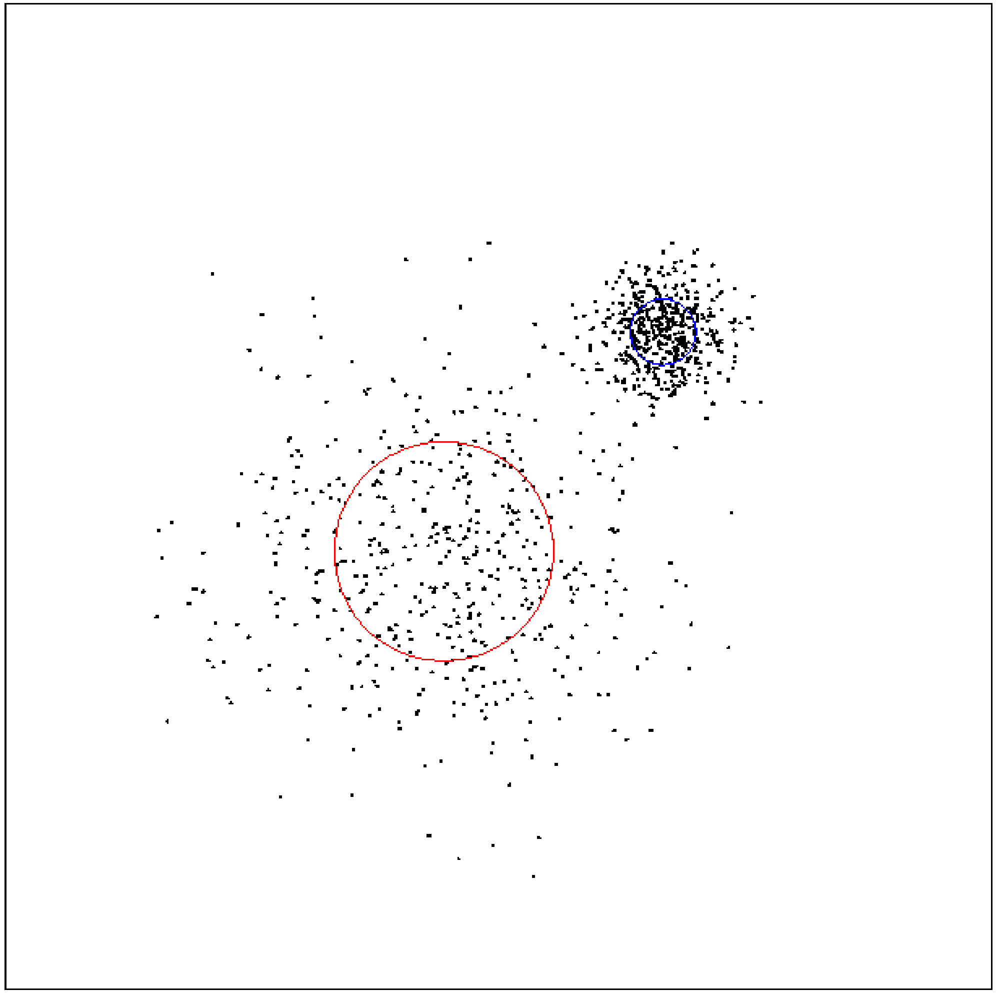
$$\mathcal{L}(y_i) = p(y_i|\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = \lambda \mathcal{N}(y_i|\mu_1, \sigma_1^2) + (1-\lambda)\mathcal{N}(y_i|\mu_2, \sigma_2^2)$$

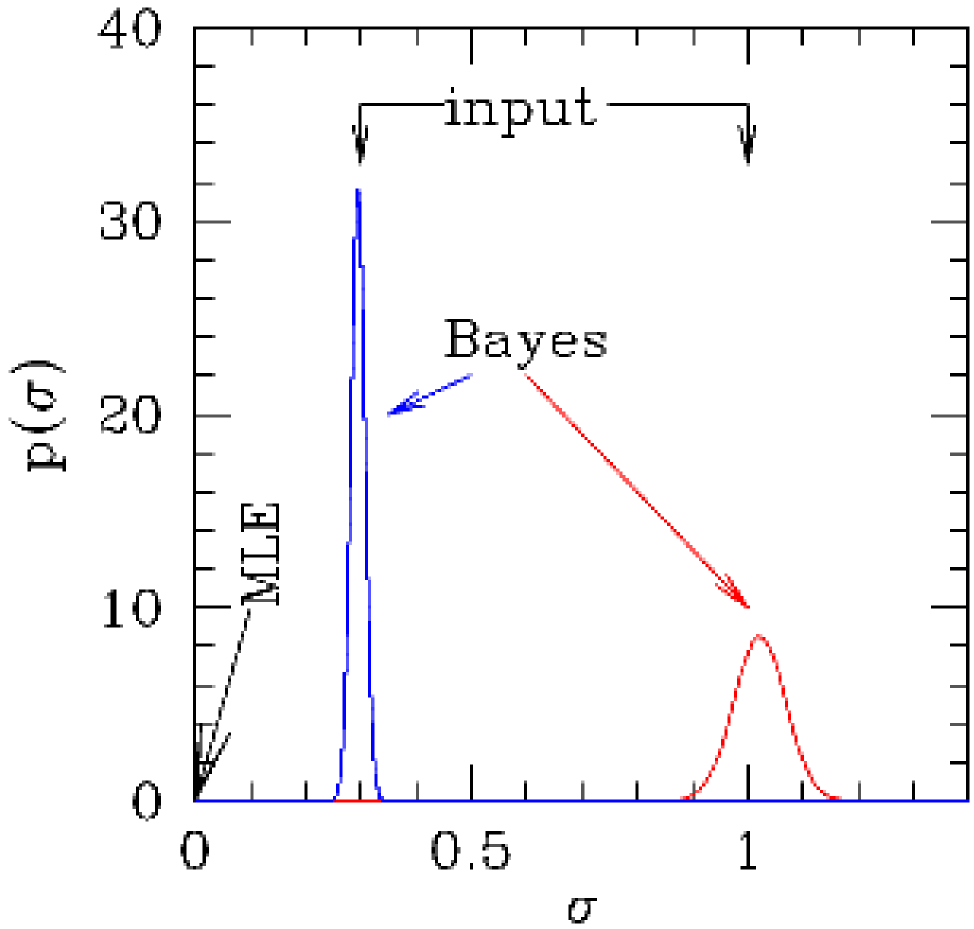
$$\mathcal{N}(y_i|\mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}}$$

$$\mathcal{L}(y) = p(y|\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = \prod_i p(y_i|\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = \prod_i \mathcal{L}(y_i)$$

$$\mathcal{L}(y) \rightarrow \infty \qquad \text{if} \quad \sigma_j \rightarrow 0 \qquad \text{and} \quad \mu_j = y_i$$

How to be right: use Bayes





Gory details

The posterior probability of σ_1 is:

Average Likelihood, not maximum

$$p(\sigma_1|...) = \int\int\int d\mu_1 d\mu_2 d\sigma_2 p(y|\mu_1,\sigma_1,\mu_2,\sigma_2,\lambda)p(\mu_1,\mu_2,\sigma_1,\sigma_2)$$

where

$$\int\int\int\int d\mu_1 d\mu_2 d\sigma_1 d\sigma_2 p(\mu_1,\mu_2,\sigma_1,\sigma_2) = 1$$

In the example we took $p(\mu_1,\mu_2,\sigma_1,\sigma_2) = const$

because the precise choice of the prior is unimportant, and computed the integral with a computer (a MCMC).

Summary about maximum likelihood and mixtures

The problem is a general characteristic of mixture of distributions with variable variance (width), not only of gaussians

The problem do not disappear 'in the long run', at the contrary the chance of failure increases with sample size! and anyway the TAC will not give you infinite time (and data).

Don't maximize the likelihood, then, for mixture distributions if you want to measure

- velocity (colour or whatever) dispersion in presence of a background
- signal intensity or surface brightness in presence of a background
- red and blue galaxies
- two objects in the field of view of my instrument
- galaxy SED (two stellar populations)

Use bayes, instead (see a later slide for an example).

A second failure of usual methods: with small samples standard results depend on the chosen parameterization. Do faint red galaxies evolve?

Aim: measure a (number) deficit of faint red galaxies as a function of redshift

De Lucia et al. (2007) compute L/F, number of bright galaxies over number of faint galaxies

Stott et al. (2007) compute F/L, i.e. the reciprocal of the above

As astronomer, I hope to see the trend on L/F mirrored on F/L, right?

Bad discovery: No, it doesn't. Why? Lets go to the heart of the problem: error weighted averages of two data points.

How to be wrong, I.

Two data points, with errors:

$(L/F)_1=3\pm0.9$ $(L/F)_2=0.3333\pm0.1$

formula for error-weighted average:

wtd average = $\Sigma\,(\mathrm{values}_i/\sigma^2)\;/\;\Sigma\,(1/\sigma^2)$

average of $(L/F)_1$, $(L/F)_2$-> 0.37, near to $(L/F)_2$ (smaller error)

Consider reciprocal values $(F/L)_1=1/(L/F)_1$ and $(F/L)_2=1/(L/F)_2$:

$(F/L)_1=0.3333\pm0.1$ $(F/L)_2=3\pm0.9$

wtd average of them -> 0.37, near $(F/L)_1$, not near $(L/F)_2$!

with reciprocal (to go back to the original quantity): 2.77 vs 0.37

(and they differ more than their error)

Further example, with points 1 σ away each other:

$(L/F)_1$=1.23±0.9 $(L/F)_2$=0.3333±0.1

av of $(L/F)_1$, $(L/F)_2$-> 0.34, very near to $(L/F)_2$ (smaller error)

the error wtd average of the reciprical: 1.48

reciprocal of the above: 0.68 midway between $(L/F)_1$ and $(L/F)_2$

What is wrong?

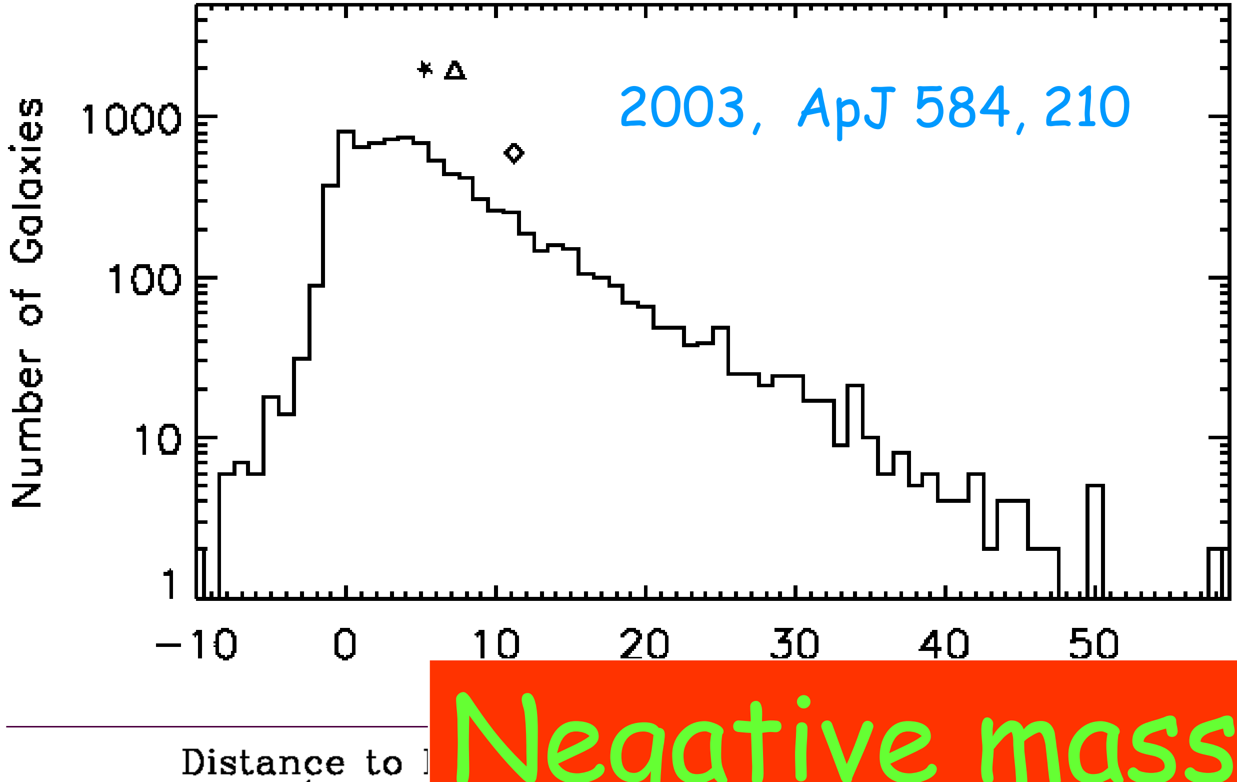
The problem is still there with a small/normal size sample.

Were you aware of the problem? Are you happy if your science conclusions depend on something about which you ignore the existence ? Have authors checked whether their fit is insensitive to the chose parameterization?

The problem comes from a) the freedom (in the frequentist paradigm) of choosing the parameterization, b) the smallness of the sample (rules usually hold asymptotically) and c) the refuse of admit that two contradictory priors have been taken (because priors does not exist for frequentists!). Bayesian methods don't have estimators among which to choose, offer methods that hold from n=0 and have priors (that should change coherently with re-parameterization).

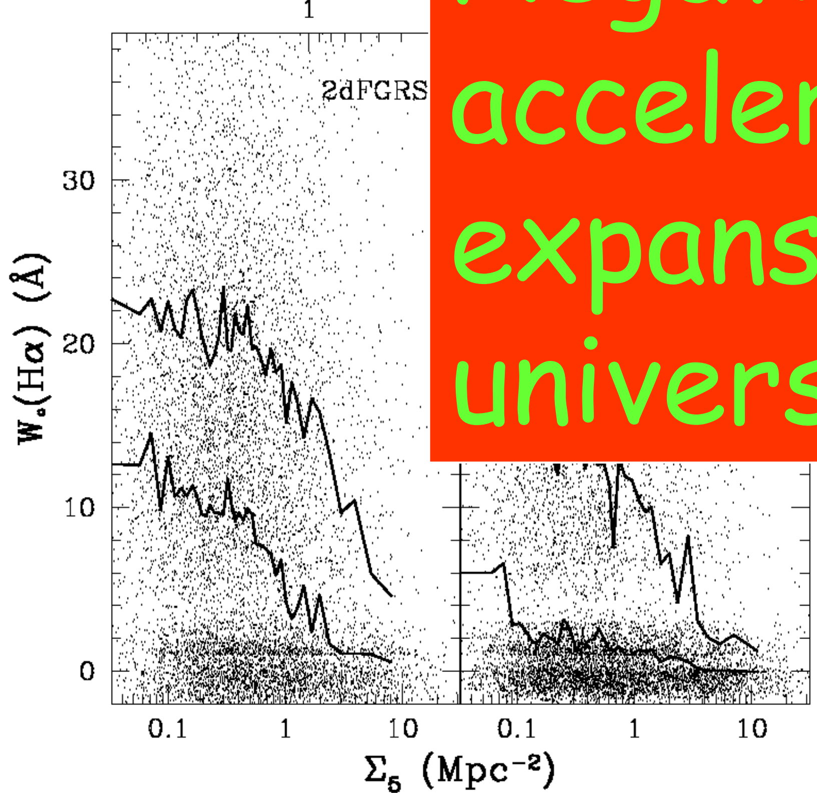
Problems don't come from small samples only:

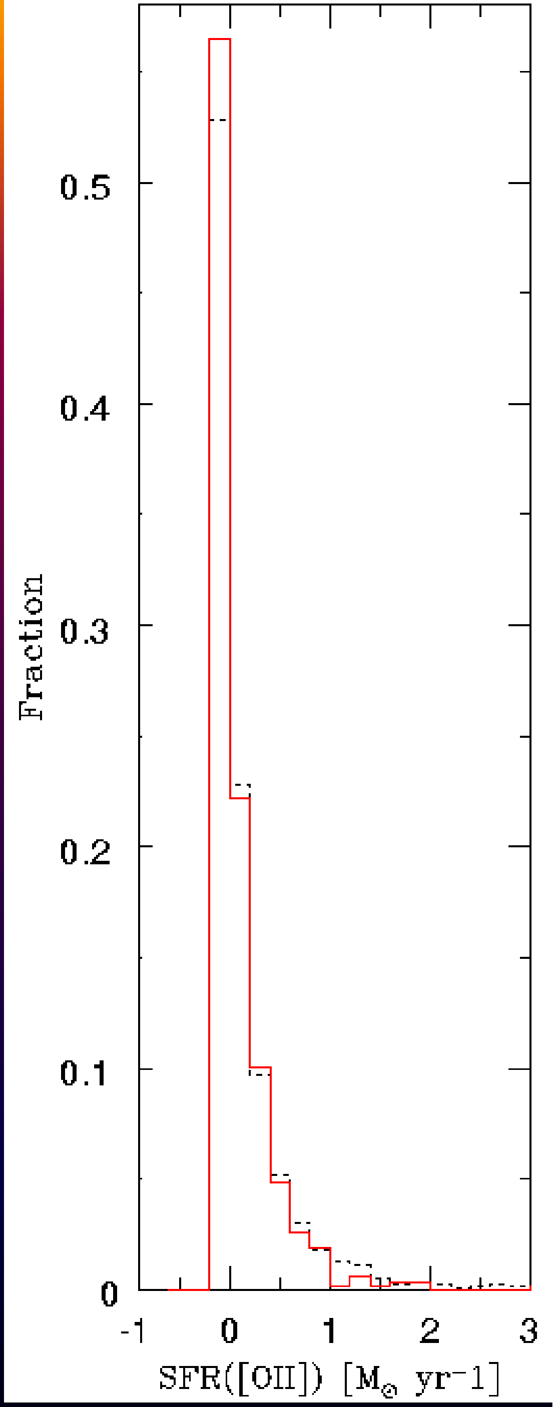
how to find impossible (unphysical) results
i.e. finding something that cannot occur,
like negative masses



2003, *ApJ* 584, 210

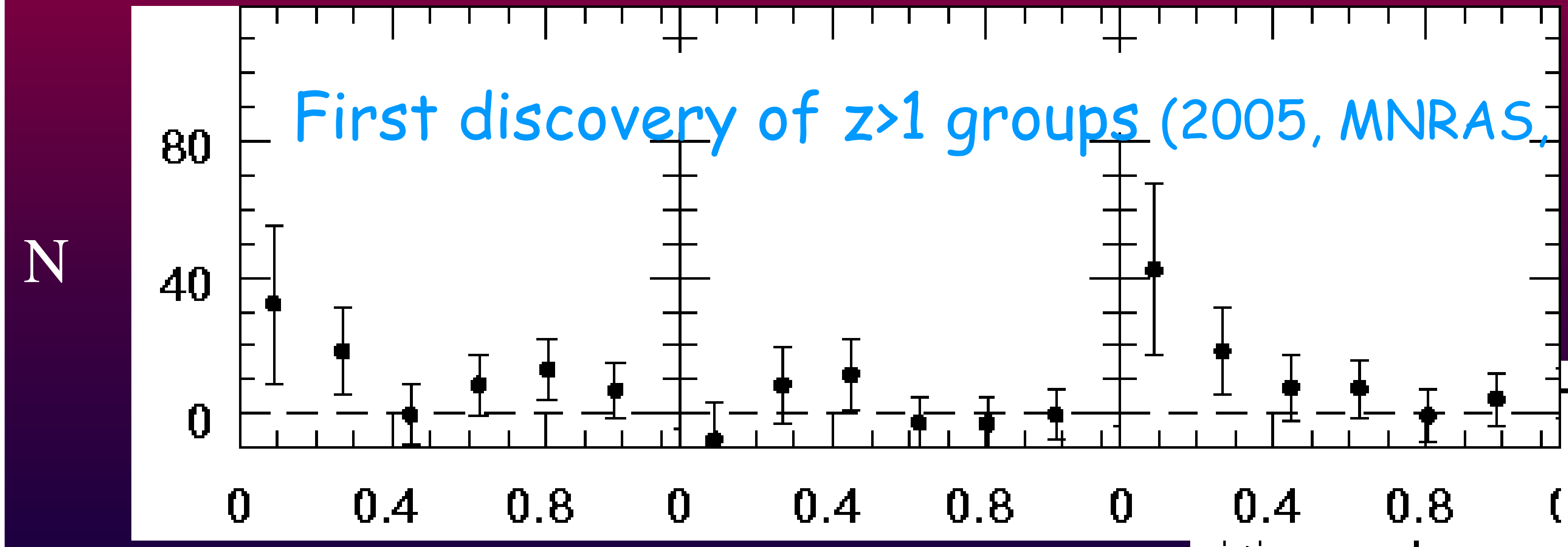
Negative masses accelerating the expansion of the universe? (joke)



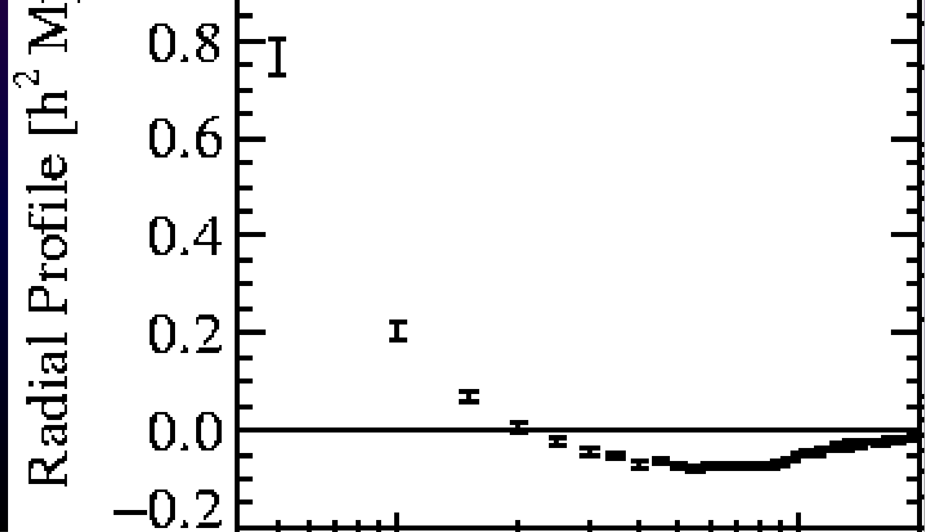


2005, *ApJ* 624, 571

Galaxies come in positive units. Maybe.



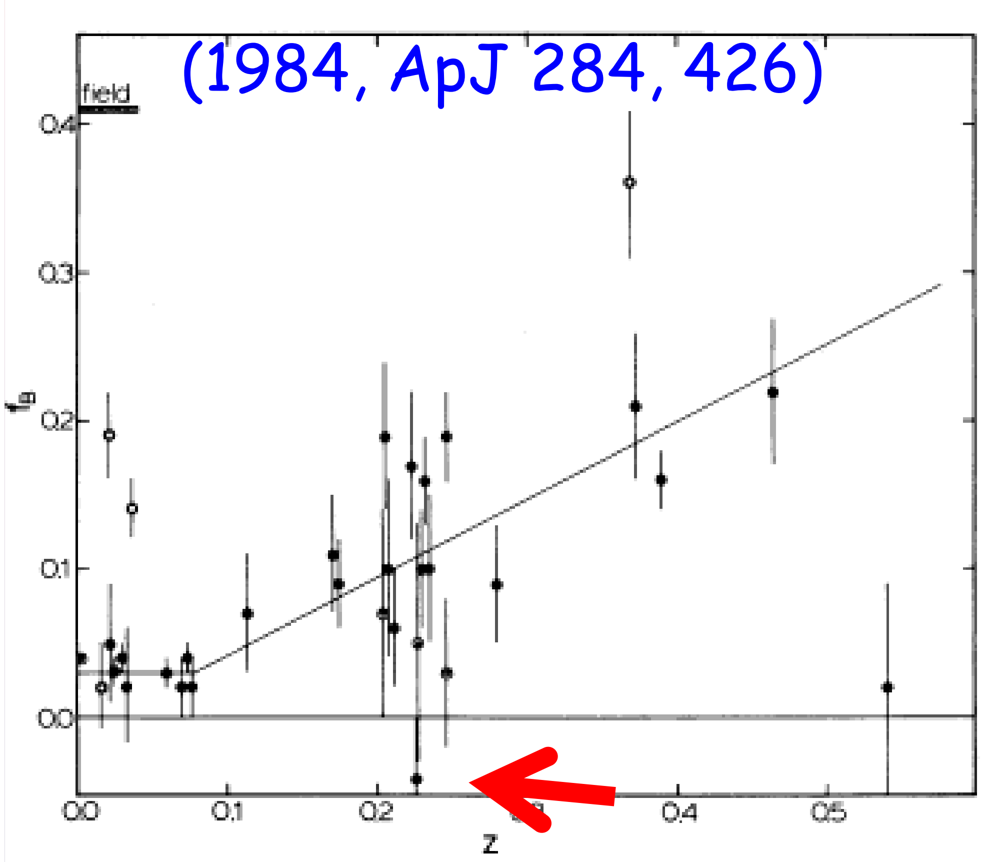
First discovery of z>1 groups (2005, MNRAS, 357, 1357)



Average of many local groups

(2005, ApJ 633, 122)

… fractions smaller than 0


(1984, ApJ 284, 426)

Maximum likelihood (best) estimates sometime fail to provide acceptable results.

Every physically acceptable value is better than the claimed best value!

What "best" does it means?

Other funnies

- fraction larger than 1
- completeness larger than 100 %
- V/Vmax >1
- Negative star formation rates
- scatter plausibly negative
- negative S/N
- imaginary (as sqrt(-300) km/s) velocity dispersions

Where the problem comes from:

Measurements near boundaries (fractions, completeness, hardness ratios, sources/features with few counts ...)

-> forgetting to account for boundaries

native instead with the Bayesian framework, that explicitly account for the fact that some values of the physical quantity are impossible.

Simple example:

p(signal|data)=c*p(data|signal)*p(signal) <- Bayes theorem

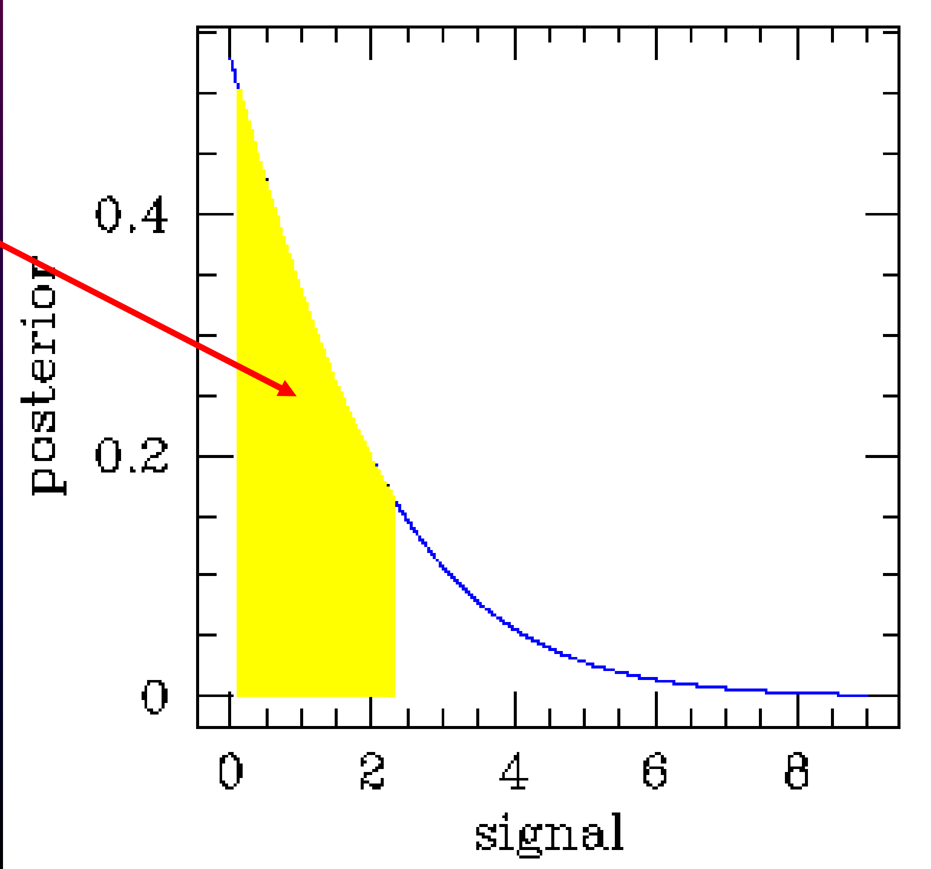
if data=3 and expected (average) backgound is 5:

= c* Poisson(x=3; λ=5+signal) * p(signal)

Since prior p(signal)=0 if signal<0, then posterior($\forall$ signal<0)=0

uniform prior, mean: 1.63; shading: shortest 68 % confidence interval

vs 3-5=-2



Malmquist/Eddington correction

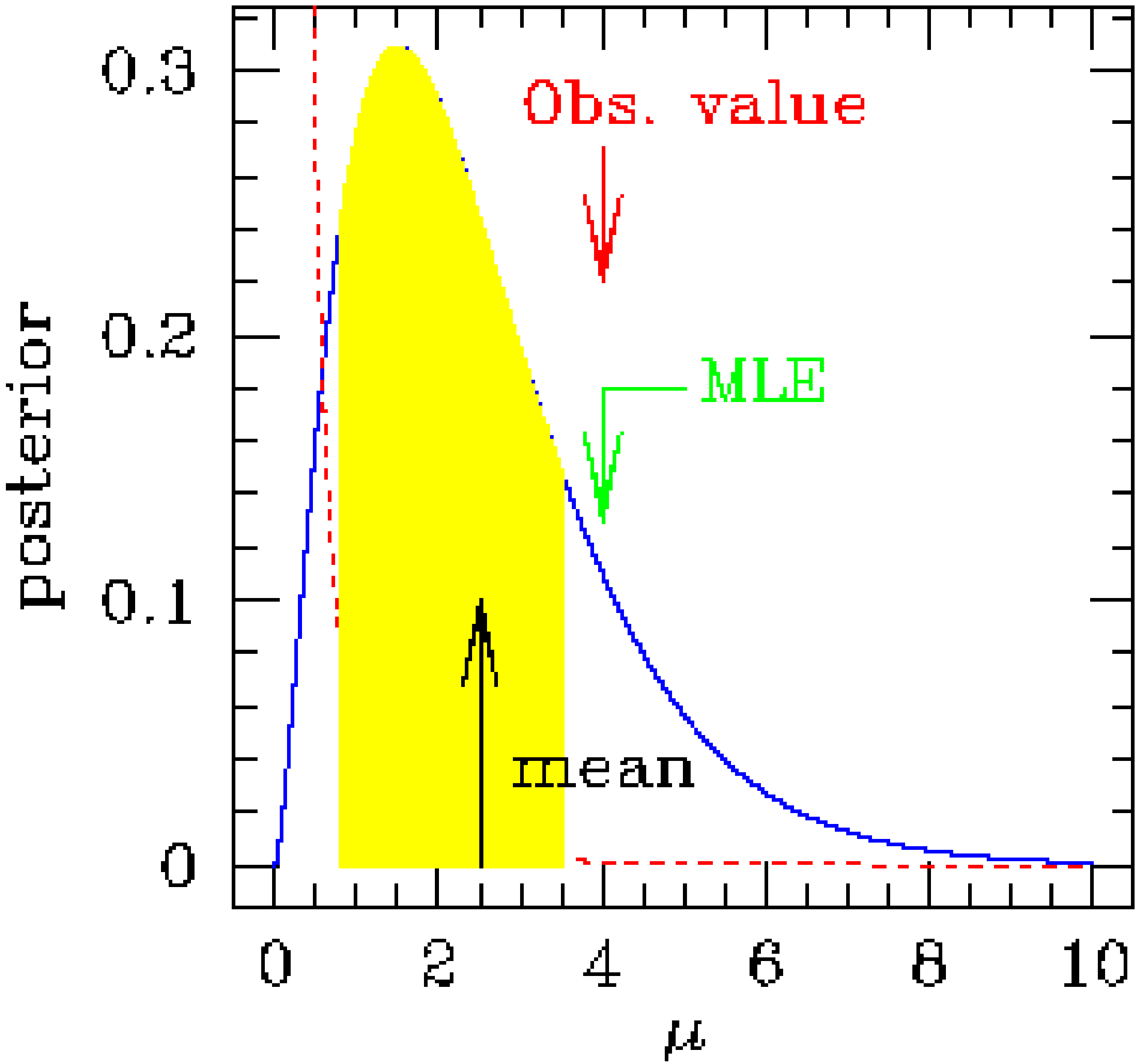
p(θ|data)= c * p(data|θ) * p(θ)

ex. modelled around Kenter et al 2005, ApJS 161, 9 (x-ray survey with sources as faint as 2-4 photons).

$p(\mu|4) = c^* \, p(4|\, \mu)\, p(\mu)$

at the studied fluxes, the prior $p(\mu)$ (=number counts for astronomers) is well known, $p(\mu)= \mu^{-\beta}$ with beta approx 2.5 (euclidian slope).

4 photons are observed but the posteriori mean is about 2.5! Well accepted by astronomers, that known that ….



Actually, the example shown that prior knowledge (the fact that there are much more faint systems than bright ones), cannot be ignored, if one does not whish be wrong most of the times (claiming that the true source flux is 4 photons). To say, ‘I have no prior’ or ‘I dislike priors’ lead to be wrong more often than claimed.

The same mis-estimation holds true for every noisy variable, such as velocity dispersion when measured with small samples, when the number density of systems changes a lot over the range where the likelihood is considerably not null.

The only way to make inferences and not to be wrong is following the Bayes theorem, any other calculus is either equivalent or wrong (Cox 1946, Am. J. Phys. 14, 1).

A selected application of mixture modeling: measuring a dispersion

$$\mathcal{L}(v) = p(v|I(v)) = \prod_i \mathcal{L}(v_i) = \prod_i \omega(v_i) I(v_i|\theta, \psi) \ \ e^{-\int_\Omega \omega(v) I(v|\theta,\psi)}$$

$$I(v) = N_{clus} \mathcal{N}(v - v_{clus}, \sigma_v^2 + \sigma_{clus}^2) + \frac{N_{bkg}}{\Delta v}$$

Gaussian +Constant

Assumed uniform priors, in next slides, because a smarter choice is not compulsory (don't change results).

Let's start simple

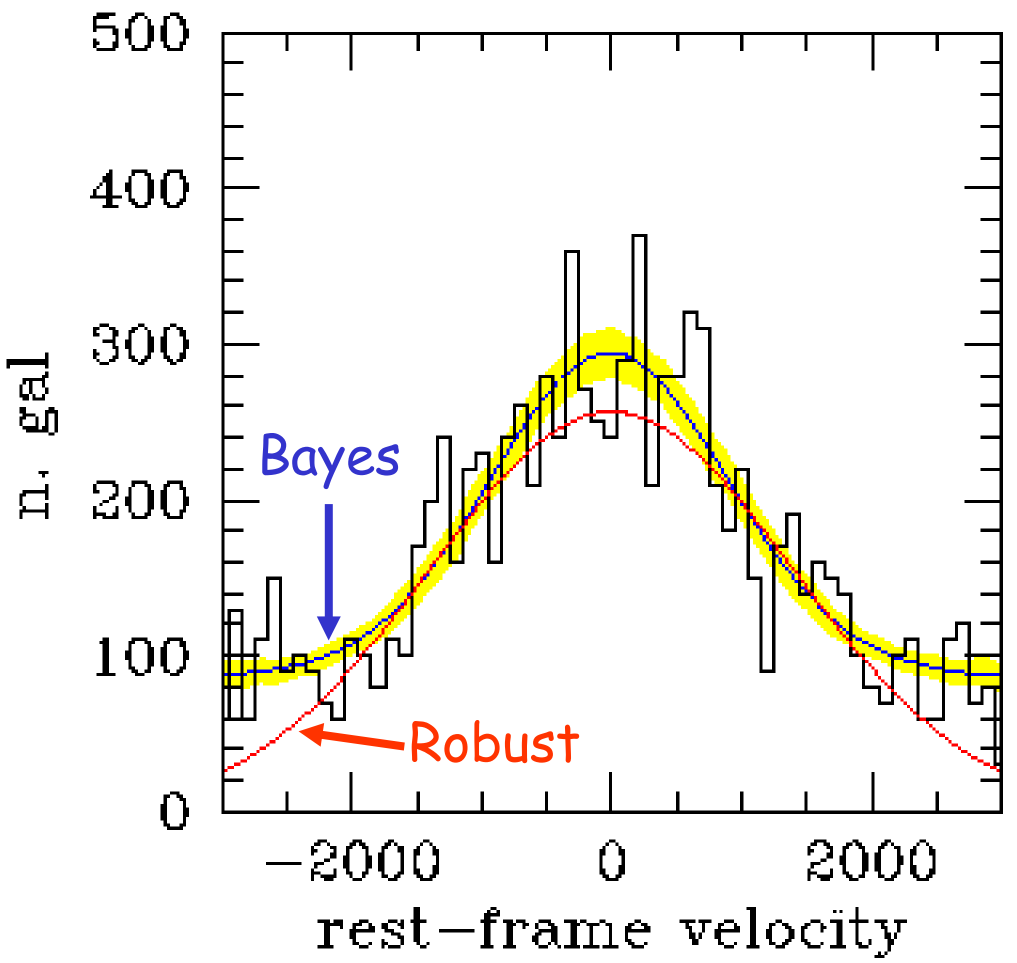
- a huge sample (1000 galaxies, 50% background)

- drawn from the very same distribution used from the analysis: Gaussian+constant

Input: v=1000 km/s

Robust (Beers): v=1400 km/s (and a negligible error)

Bayes: v=940±85

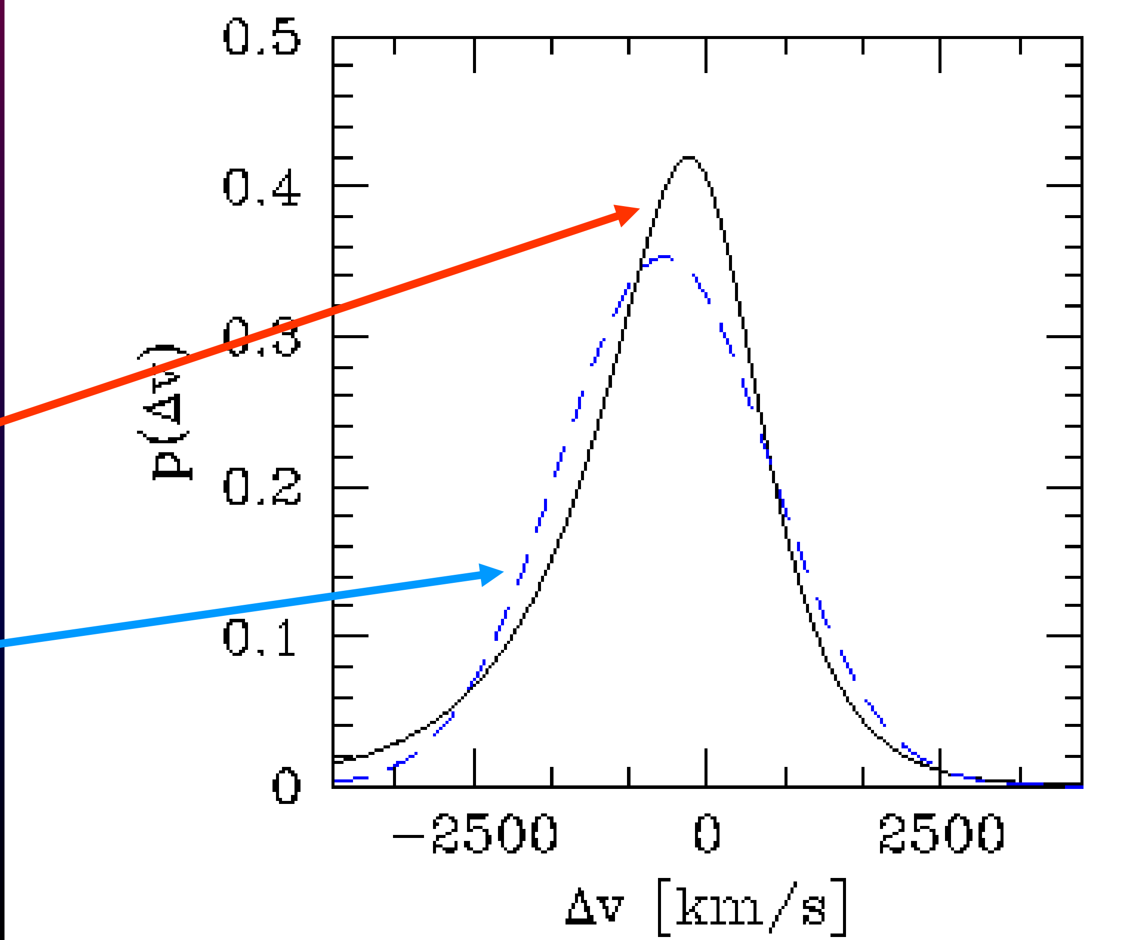


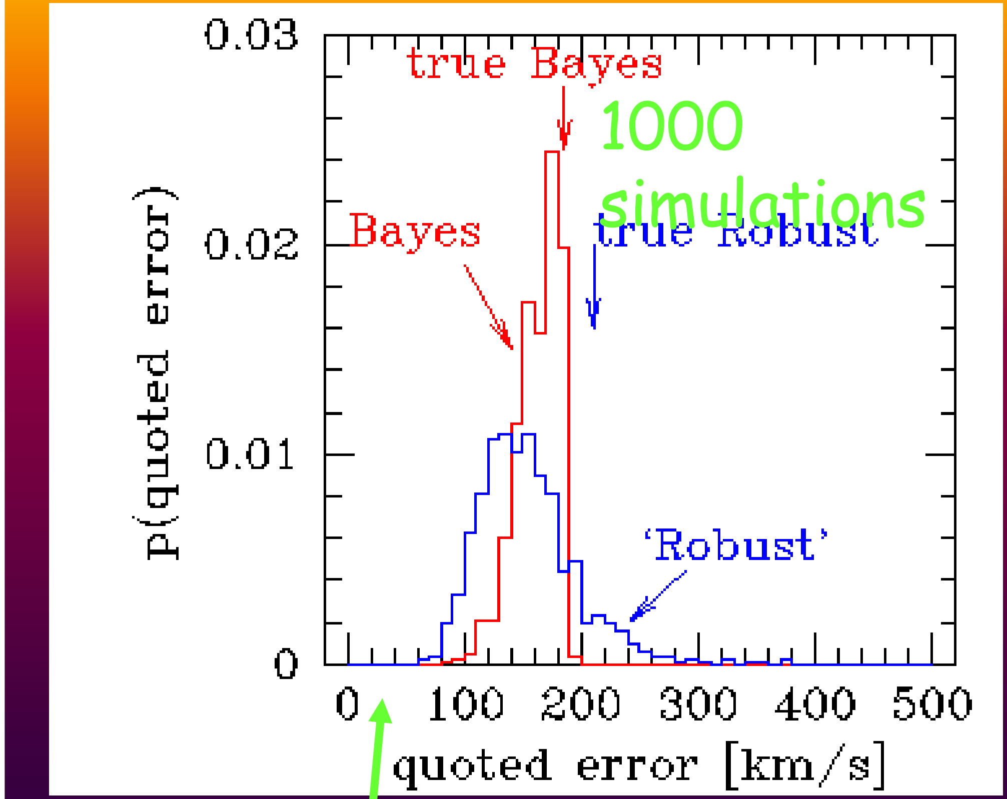
Making things difficult, i.e. nearer to reality

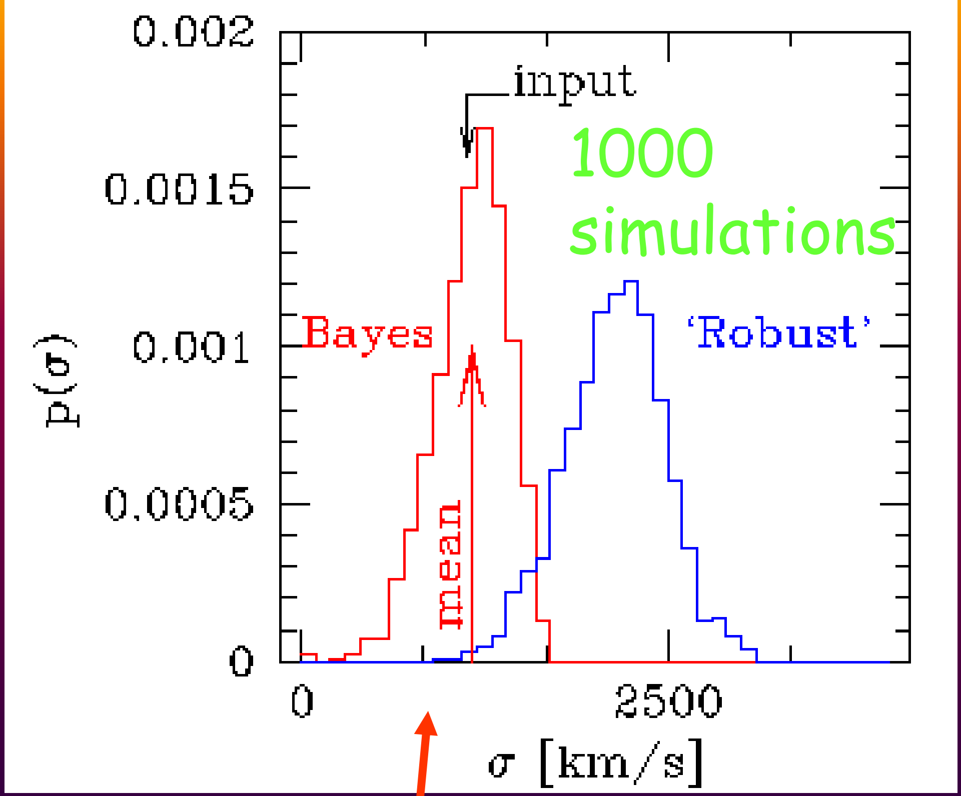
In real world applications we ignore if our data are drawn from the supposed distribution. Here: do my galaxies come from a Gaussian or something similar to it?

Added difficulty: simulated data from a perturbed (tilted) distribution and Bayes-analysed supposing it is Gaussian

Summary: as before …







25 galaxies, no bkg

25 galaxies, 50% bkg

The same holds true for synthetic data generated from a Gaussian.

Science: all x-ray dark cluster for their become normal x-ray emitting clusters, because of the 'Robust' over-estimation of σ_v or the neglected Bayes (Malmquist) correction (Andreon et al. 2007, MNRAS in press, arXiv:0710.1023)

Reminder

The developed mixture model of inhomogeneous (of varying amplitude) Poisson process holds for whatever quantity subject to a Poisson fluctuations, and thus it is applicable to other problems. Actually, I develop it to measure the (X-ray) flux and core radius of a very faint extended source (a cluster of galaxies): it is just matter of renaming the variable v with the name (ra,dec), and replace the Gaussian in I(v) with a β-model(ra, dec, ellipticity, core radius, β).

Regressions

Astronomers are interested in estimating as quantities vary as a function of each other: e.g. Tully-Fisher, Faber-Jackson, Magorrian relations, Fundamental-Plane, cluster scaling relations, GRBs (Ghirlanda) relations, etc.

How to determine the parameter of these regressions? There is no consensus in the astronomical literature: direct-, inverse-, orthogonal-, Bivariate Correlated Error and Intrinsic Scatter- (BCES), Measurement Error and Intrinsic Scatter- (MEIS), etc.   -fit?

There is no consensus among astronomers, but there is just one single solution among Bayesian statisticians.

For variables with names appropriate for the colour-magnitude relation (assumed linear, with an intrinsic scatter σ_{intr}) and data with hetheroscedastic (not of equal magnitude) errors, the likelihood is given by a Gaussian:

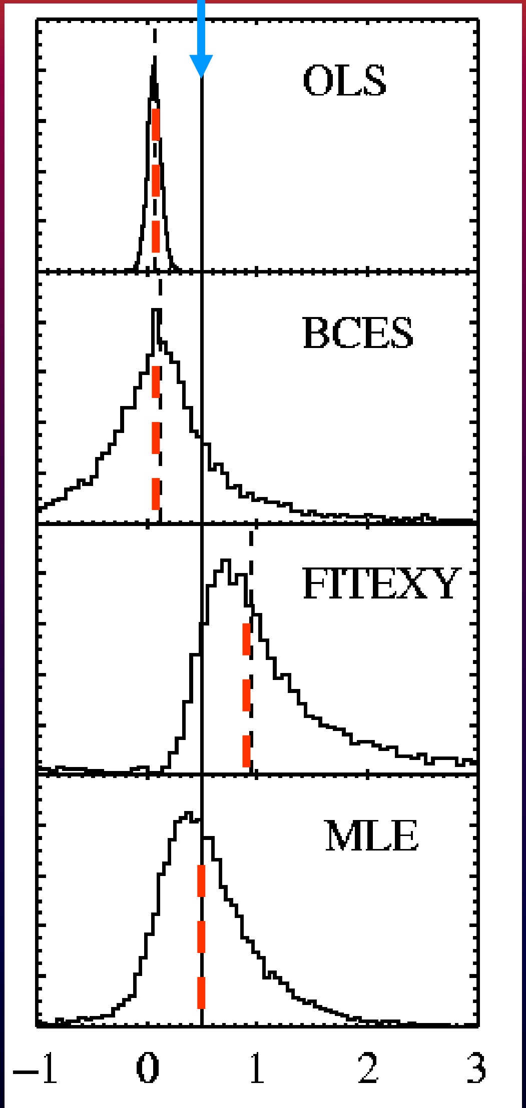
$$\mathcal{L}(m_i, col_i) = p(m_i, col_i | a, c, \sigma_{intr}, nobkg) \propto \mathcal{N}(\mu_{col_i} - a\, m_i - c, \sigma^2_{intr} + \sigma^2_{\mathrm{col}_i} + a^2 \sigma^2_{m_i})$$

e.g. D'Agostini (2003) and Gelman et al (2004) books, or Andreon (2006, MNRAS 369, 969), Kelly et al. (2007, ApJ 665, 1489).

Kelly et al 2007, ApJ 665, 1489

Why astronomer' techniques do not work

Input value



Estimated slope

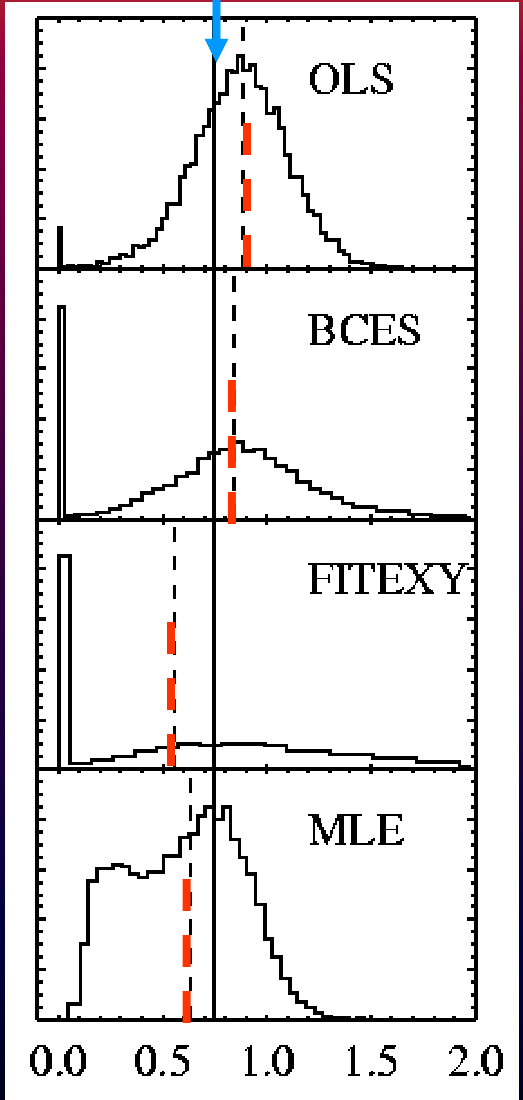
Ordinary Last Square fit

Bivariate Correlated Error and Intrinsic Scatter fit

Press et al. (numerical recipes)

Simplified bayesian solution

Input value



Estimated intrinsic dispersion

As usual, reality is more complicated: there are interlopers/outliers!

Use then a mixture of regression:

$$\mathcal{L}_i = p(m_i, col_i | a, c, \sigma_{intr}, \alpha, M^* \phi^*) = \Omega_j bkg(m, col) + \\ + \delta_c \Omega_j \mathcal{N}(\mu_{col_i} - a\, m_i - c, \sigma^2_{intr} + \sigma^2_{col_i} + a^2 \sigma^2_{m_i}) Schechter(m_i | \alpha, M^* \phi^*)$$

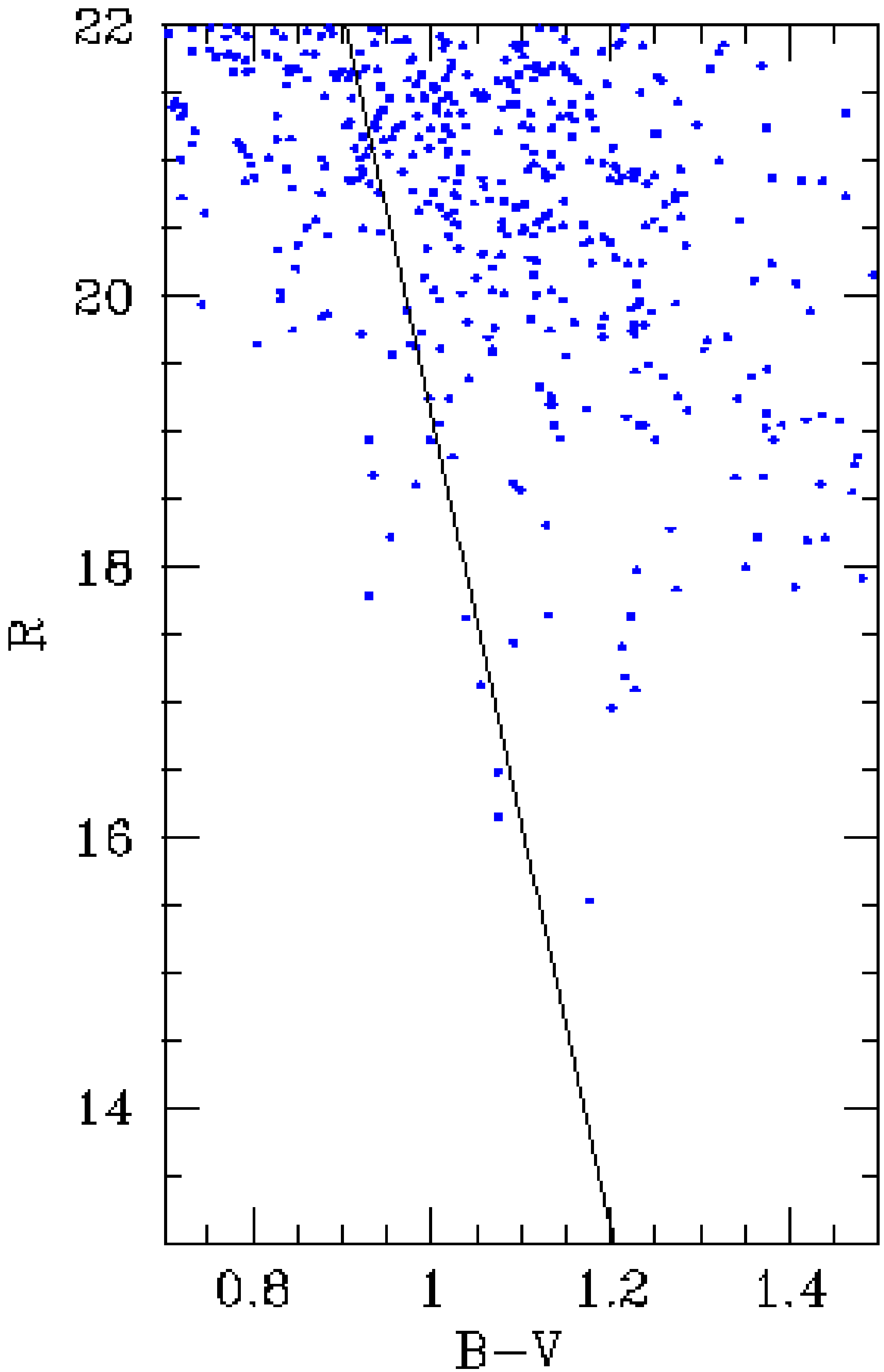
As usual for iid data:

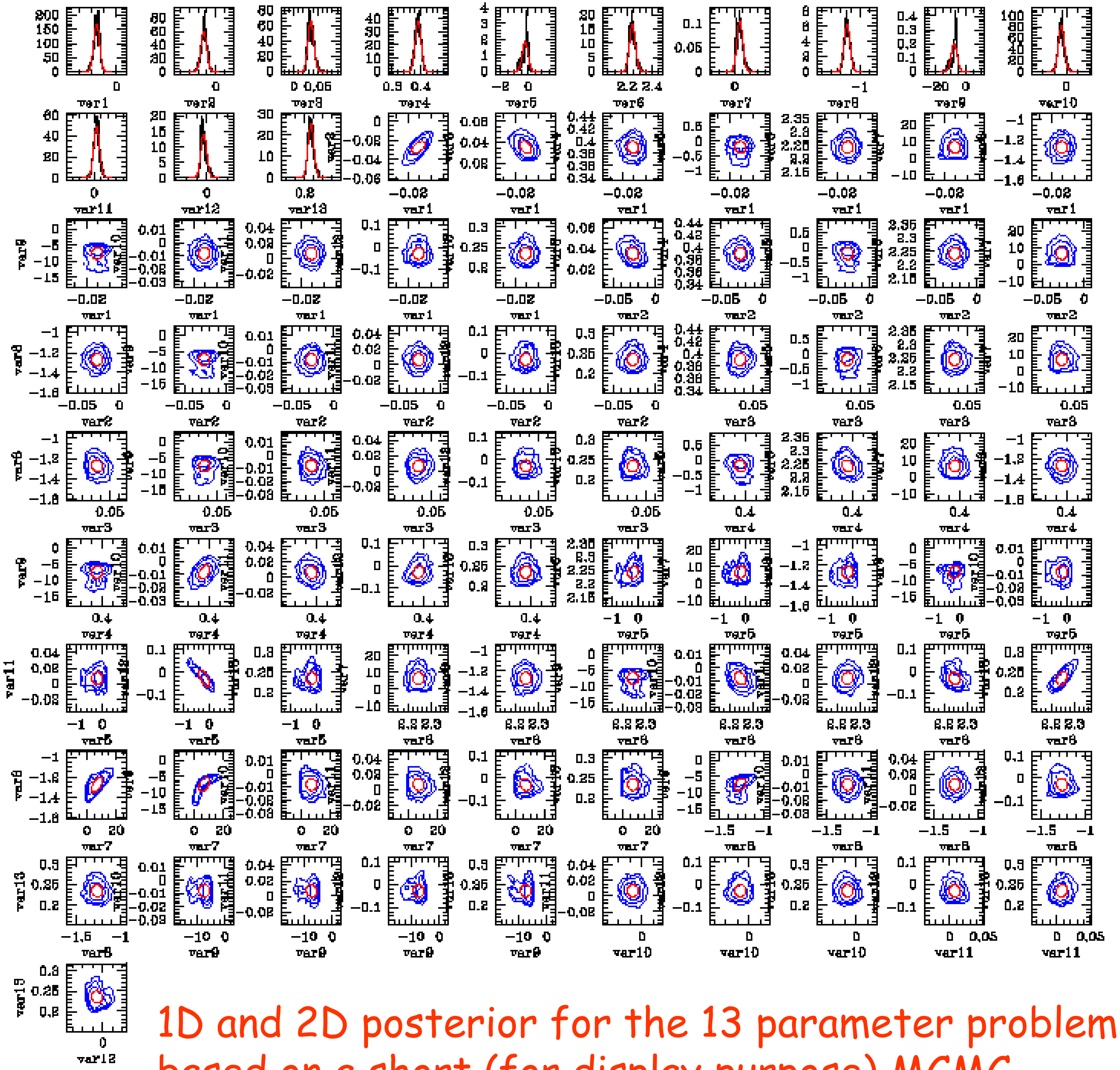
$$\mathcal{L} = \prod_i \mathcal{L}_i$$

Astro: the scatter around the colour-magnitude relation put a strong constraint on the age of stars in these objects and, indirectly, on the ages of the galaxies themselves.

Problem: background galaxies (about 4 for every cluster galaxy).

Abell 1185 cluster (Andreon, Cuillandre et al. 2006, MNRAS 372, 60)





1D and 2D posterior for the 13 parameter problem, based on a short (for display purpose) MCMC

Summary

Bayesian methods work when we mostly need them:

- when other methods returns unsatisfactory results, as in the case of the toy mixture model, of velocity dispersion, or when unphysical values are found
- we don't need to await a larger sample
- when there is no agreement among astronomers on the good procedure to use, as for the regression problem
- embody the Malmquist or Eddington correction
- (clarify what other procedures compute, as p-values or weighted average of small samples; p-values to be shown if time allow)

Don't forget that the scientific paradigm suggests to prefer a procedure that work over one that might, or might not, work.

Addressed parameter estimation problems:

- fractions, completeness, hardness ratios
- scale (e.g. velocity dispersion), with heteroscedastic errors when either a summary is available (a Beer scale) or when the full information is available (individual velocities), accounting for instrument (limited) resolution
- luminosity function
- spatially structured Poisson process (x-ray flux of an extended source, richness)
- trends (linear regressions) between quantities (e.g. CM, …), each with heteroscedastic errors and in presence of an intrinsic dispersion

all in presence of a background.

see also my bayesian primer for astronomers from my homepage

Model selection problems addressed:

- Is the LF Schechter-like or should a dip be added ?
- Is the CM linear or bended?
- Which models describe the mass buildup of galaxies in cluster, and which models should be rejected?
- Do my (mass buildup) model need to be refined (with a further degree of complexity) in light of the data?
- Does the fraction of obscured AGN depend on flux, luminosity, or redshift?
- Is the dependence of the fraction of obscured and absorbed AGN vs my favorite quantity the same?
- Does the fraction of blue galaxies depend on the cluster velocity dispersion?

see my Inference homepage

Thank you

http://www.brera.mi.astro.it/~andreon/inference/Bayesian_primer_for_astronomers.html

http://www.brera.mi.astro.it/~andreon/inference/Inference.html

p-values are not probability of the hypothesis

the probability of rejecting the null hypothesis is often taken equal to the p-value, i.e. the number returned by tests like Kolmogorov-Smirnov, χ^2, Student's t, Wilcoxon rank, Kendal, …. tests.

1) these tests sometime return widely different numbers (such as 0.001 and 0.861). How it is possible if the (wished-for) probability takes (hopefully) one single value?

2) how it is possible that some of them do not use errors: may a probability not depend on error bars? for example do you trust a trend equally if errors on data points are huge or small (and you use a test statistics that does not use errors)?

All these tests return p(>Data |hypothesis) not p(hypothesis|Data)

Let's go with an astronomical example

Hp: "a source is there"

p-value=p(>I_0|background)= 0.003 = 0.3 %

Is my target real at 99.7 %? (=1 minus the p-value)

x is the fraction of the sky occupied by sources

In a sample of N sky positions, there are xN sources and 0.003(1-x)*N (false) detections. Therefore, there are xN true source out of the total xN+ 0.003(1-x)*N detections. The fraction of true sources is the ratio of the above numbers:

p(true source | I_0, x)= x/(x+ 0.003(1-x))

1) if x=0.07 % (appropriate for Chandra typical exposures)

p(true source|…)= 19 % (not 99.7 !) i.e. 4 out 5 detection are junk. "99.7" means that 3 out 1000 are junk! It's the same, in your opinion?

2) if x=5 %, p(true source|…)= 95 %; if x=20%, p(true source|…)= 99%

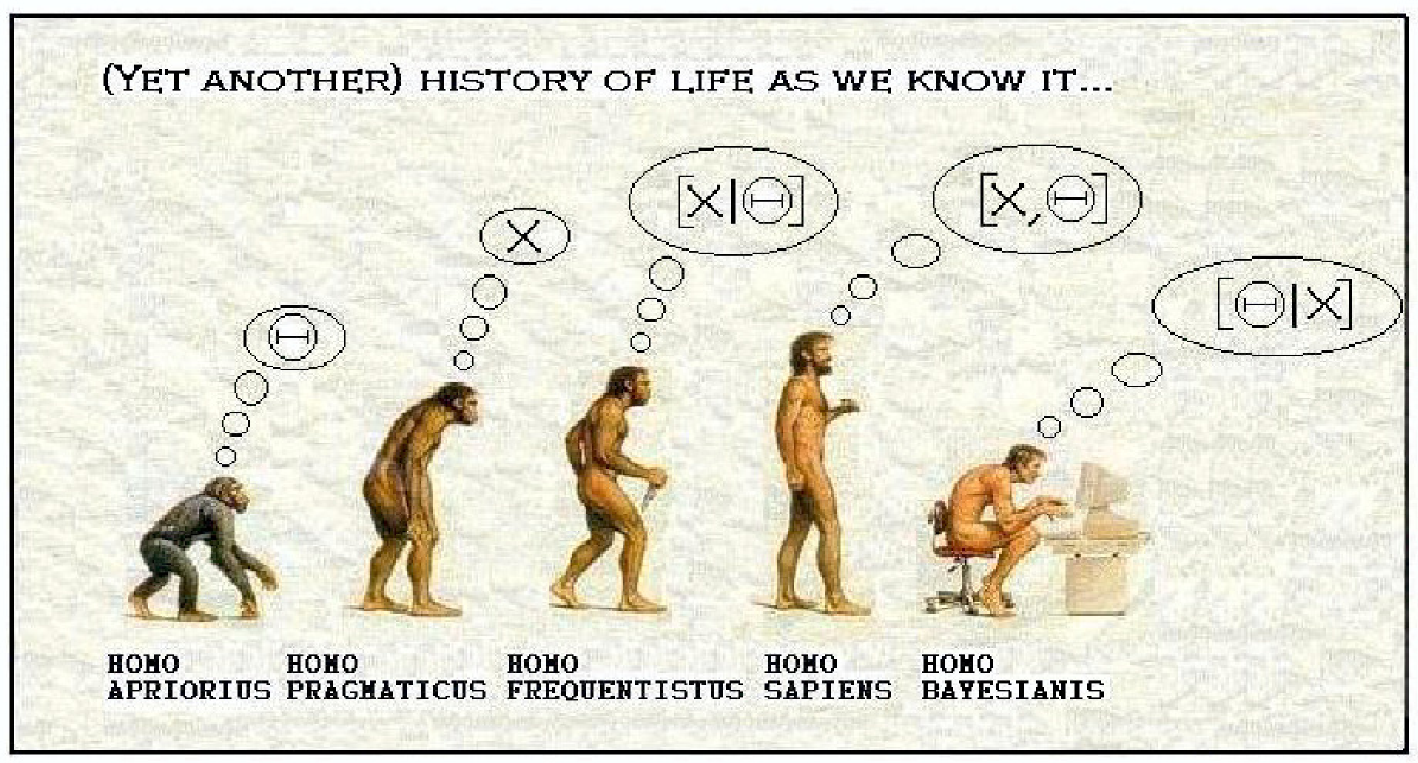
p values and probability for the hypothesis are two different things, that

takes different numbers, don't mistakenly take one for the other!

The whished-for number, the probability of the null hypothesis, depends on the prior probability (x in our example), yet almost every time a trend/effect is stated to exist and a ‘probability to reject the null hypothesis is quoted’ no prior probability is taken (stated). The prior may change a 99.7% (i.e. almost sure) result in a 19 % sure (i.e. highly uncertain) result.

Don’t quote p-value as probability of rejecting the null hypothesis, because they are not such thing.

(yet another) Chronology (from S. Andreon's web)



- Homo ***apriorius*** establishes probability of an hypothesis, no matter what data tell.
- Homo ***pragamiticus*** establishes that he is interested by the data only.
- Homo **frequentistus** measures probability of the data given the hypothesis.
- Homo ***sapiens*** measures probability of the data and of the hypothesis.
- Homo ***bayesianis*** measures probability of the hypothesis, given the data.