Statistics in Astronomy (II) applications

Shen, Shiyin

Contents

I. Linear fitting

$$Y = a\,x + b$$

Famous linear relations in astronomy

- period -luminosity relation of Cepheids
- M_{BH}-σ relation
- Tully-Fisher (L - V_{max}) relation
- Fundamental plane of ellipticals
- L-T, L-σ relation of groups and clusters
- All are statistical scaling relations, none of them are first principle like $F=ma$

Nature of the scaling relations

- Observables: (x_i, y_i) with error $(\Delta_{x,i}, \Delta_{yi})$
- First, we should find some correlations, e.g. rank analysis
- To the first order, all the correlations are linear
- $Y = a*X + b + \sigma$
  - σ is the intrinsic scatter, may not be a constant
- Observables maybe biased
  - e.g. some low-luminosity galaxies are not observed at given V_{max}
- Some observables may only be upper limits
  - E.g. we only get the upper-limit of L_x of some cluster

Ordinary Linear regression OLS(y|x)

- y_i with measurement error σ_i

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

Code: *fit* in numeric recipes

Error on both x and y

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2}$$

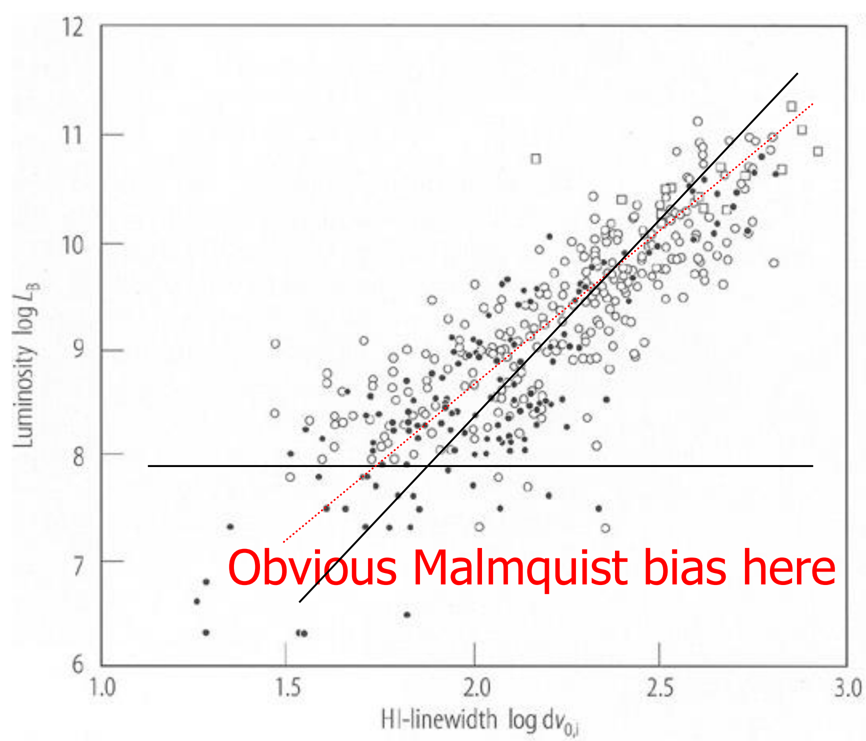
Code: *fitexy* in numeric recipes

b ~ biased to infinity

Caveat: choose proper parameterization

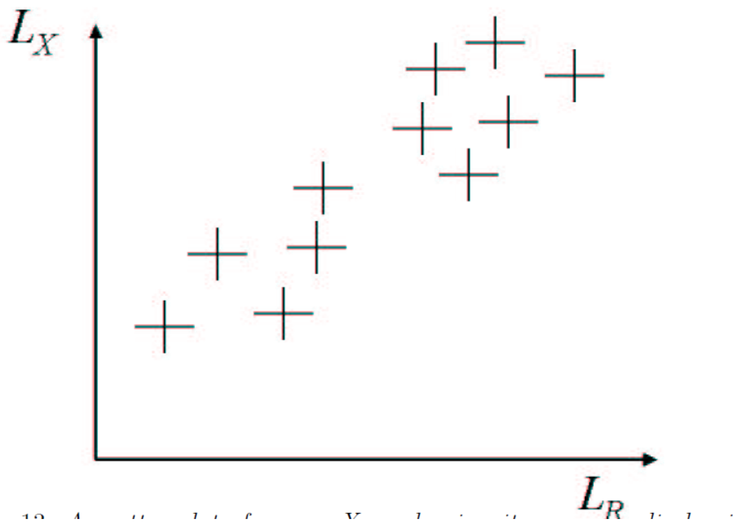
- If we fit M = *a* log W+ *b*, *a* will be biased to smaller values

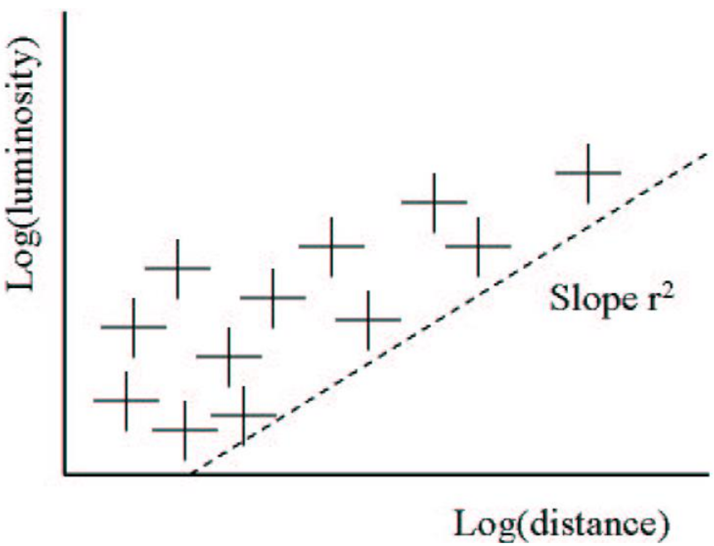
- Fit log W = a' M + b' is better
  - At given M, no obvious in W



Eddington(Malmquist) bias

- Distance dependent observable
  - Eddington (1915) Malmquist(1920)
  - In magnitude limit sample, more faint source scattered in than bright source scattered out





Six different linear regression

- Reference
  - Linear regression in astronomy I (1990, ApJ,364,104)
    - Different regression method
  - Linear regression in astronomy (1992ApJ...397...55)
    - Truncated, censored data
- IDL code: sixlin
  - Ordinary Least Squares (OLS) Y vs. X (c.f. linfit.pro)
  - Ordinary Least Squares X vs. Y
  - Ordinary Least Squares Bisector
  - Orthogonal Reduced Major Axis ;
  - Reduced Major-Axis
  - Mean ordinary Least Squares

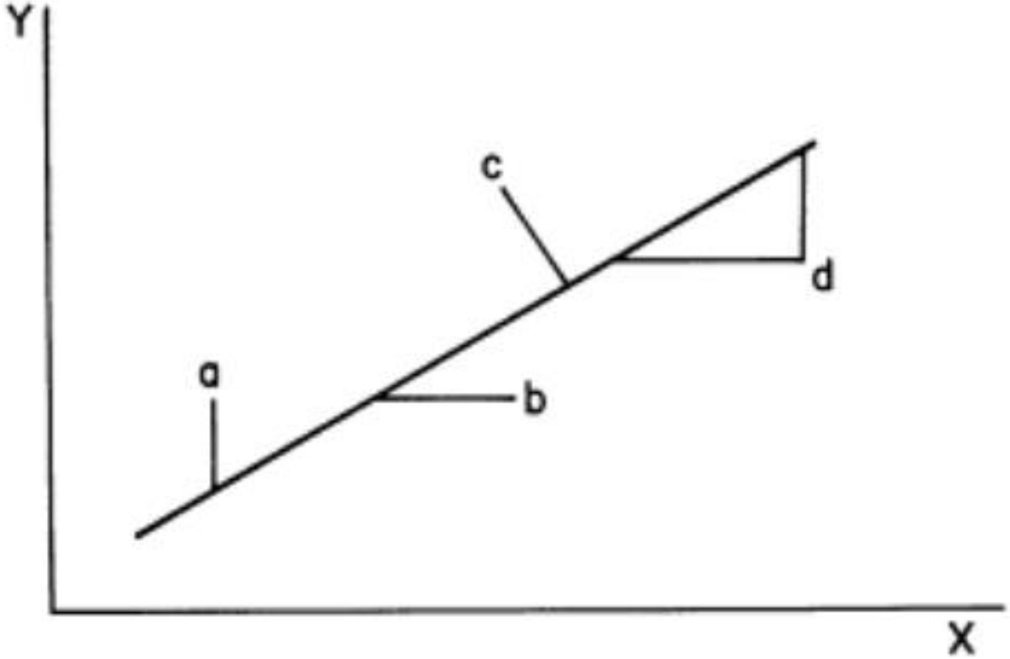


FIG. 1.—Illustration of the different methods for minimizing the distance of the data from a fitted line: (*a*) OLS($Y \mid X$), where the distance is measured vertically; (*b*) OLS($X \mid Y$), where the distance is taken horizontally; (*c*) OR, where the distance is measured vertically to the line; and (*d*) RMA, where the distances are measured both perpendicularly and horizontally. No illustration of the OLS bisector is drawn in this figure.

- The applicability of the procedures is dependent on the nature of the astronomical data under consideration and the scientific purpose of the regression.
- For problems needing symmetrical treatment of the variables, the OLS bisector performs significantly better than orthogonal or reduced major-axis regression.

Error on both x and y and with a constant intrinsic scatter σ

$$\ln L = -\frac{1}{2}\sum_i \ln\left(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2\right) - \sum_i \frac{\left[\hat{y}_i - (a\hat{x}_i + b)\right]^2}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)} + \text{constant.}$$

BCES (Akritas & Bershady, ApJ 470, 706 1996)

- Regression with correlated measurement errors and intrinsic scatter
  - allows for measurement errors on both variables
  - allows the measurement errors for the two variables to be dependent
  - allows the magnitudes of the measurement errors to depend on the measurements
- Intrinsic scatter: constant
- IDL code: BCES.pro (BCES: bivariate, correlate errors and scatter)

Linear fitting of scaling relations with intrinsic scatter

$$\ln L = -\frac{1}{2}\sum_i \ln\left(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2\right) - \sum_i \frac{\left[\hat{y}_i - (a\hat{x}_i + b)\right]^2}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)} + \text{constant.}$$

Special cases

Robust estimation

- Data with outlier

$$\text{minimize over } \mathbf{a} \qquad \sum_{i=1}^{N} \rho\left(\frac{y_i - y(x_i;\mathbf{a})}{\sigma_i}\right)$$

$$\sum_{i=1}^{N} |y_i - a - bx_i|$$

See Numeric recipes C15.7

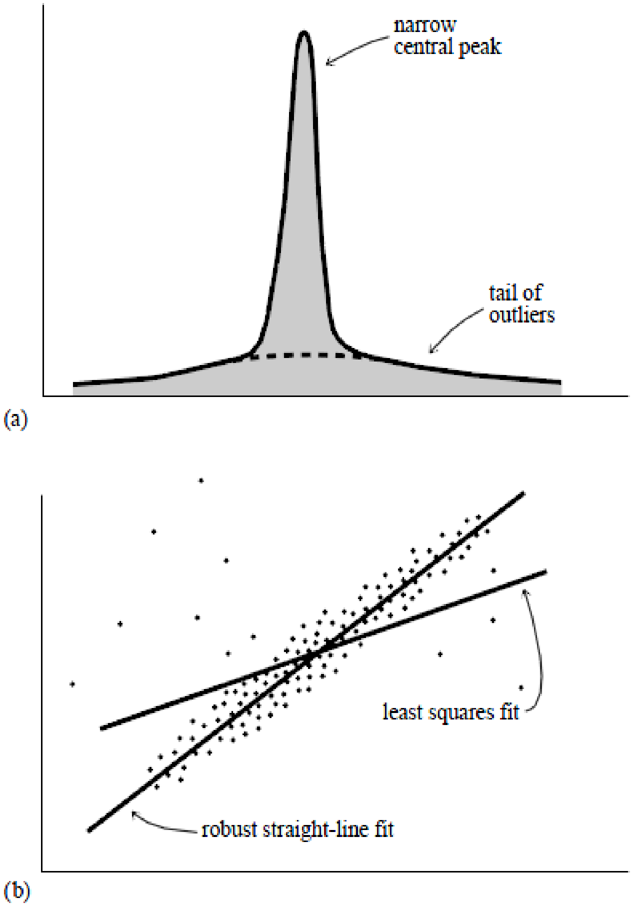
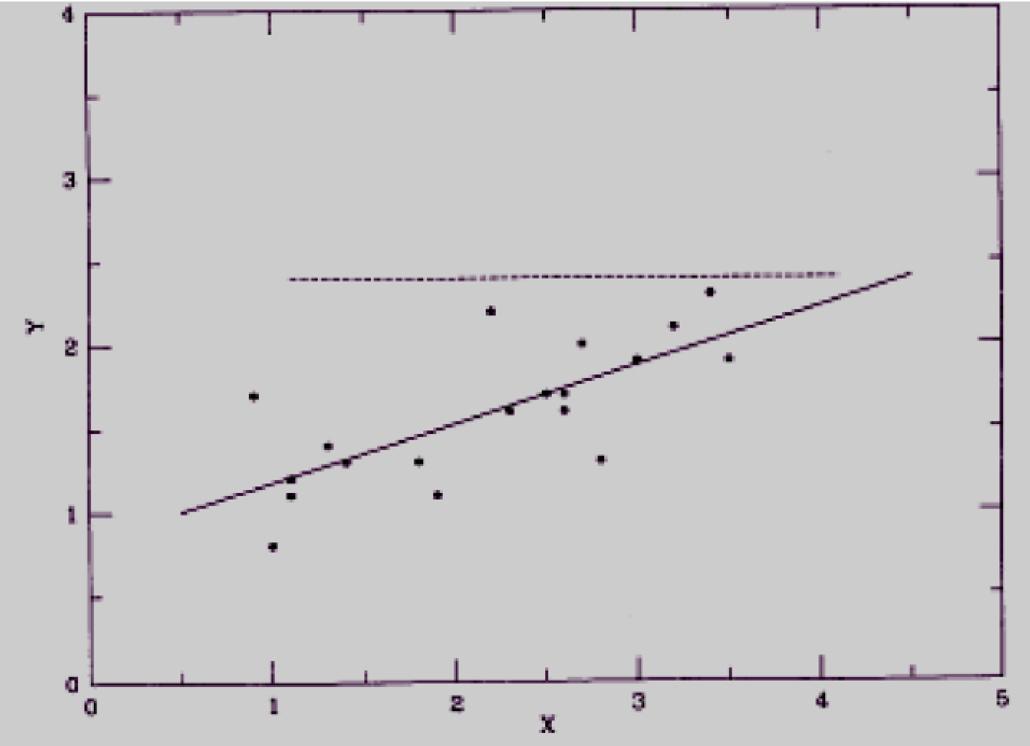


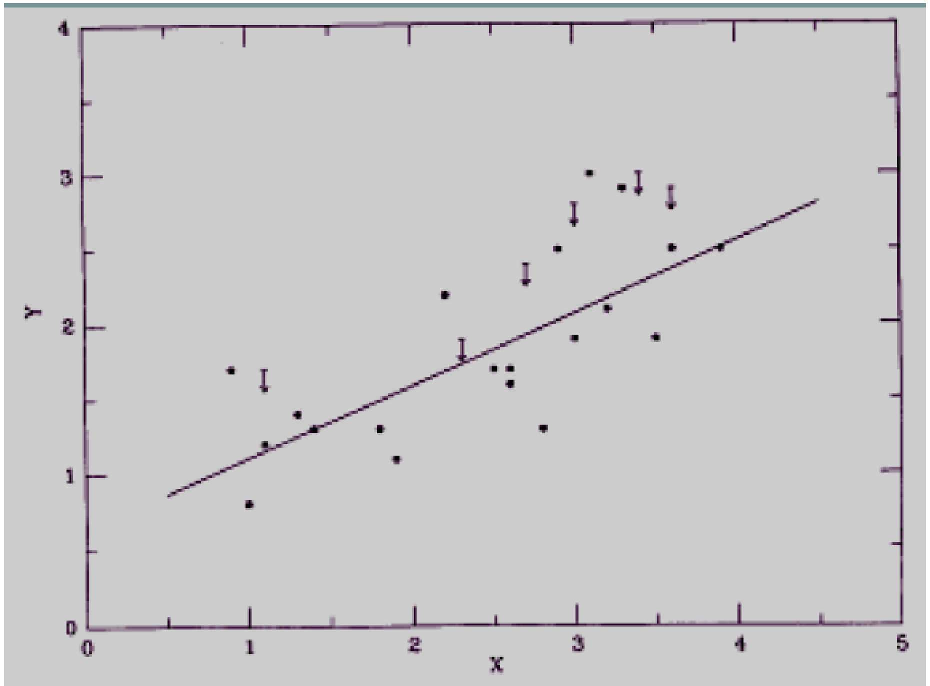
Figure 15.7.1. Examples where robust statistical methods are desirable: (a) A one-dimensional distribution with a tail of outliers; statistical fluctuations in these outliers can prevent accurate determination of the position of the central peak. (b) A distribution in two dimensions fitted to a straight line; non-robust techniques such as least-squares fitting can have undesired sensitivity to outlying points.

Truncation due to flux limits



Malmquist bias in Hubble diagram (Deeming, Vistas Astr 1968, Segal, PNAS 1975)

Censoring due to non-detections



Presented for astronomy by Isobe, Feigelson & Nelson (ApJ 1986)
Implemented in Astronomy Survival Analysis (ASURV) package

A more straight-forward way

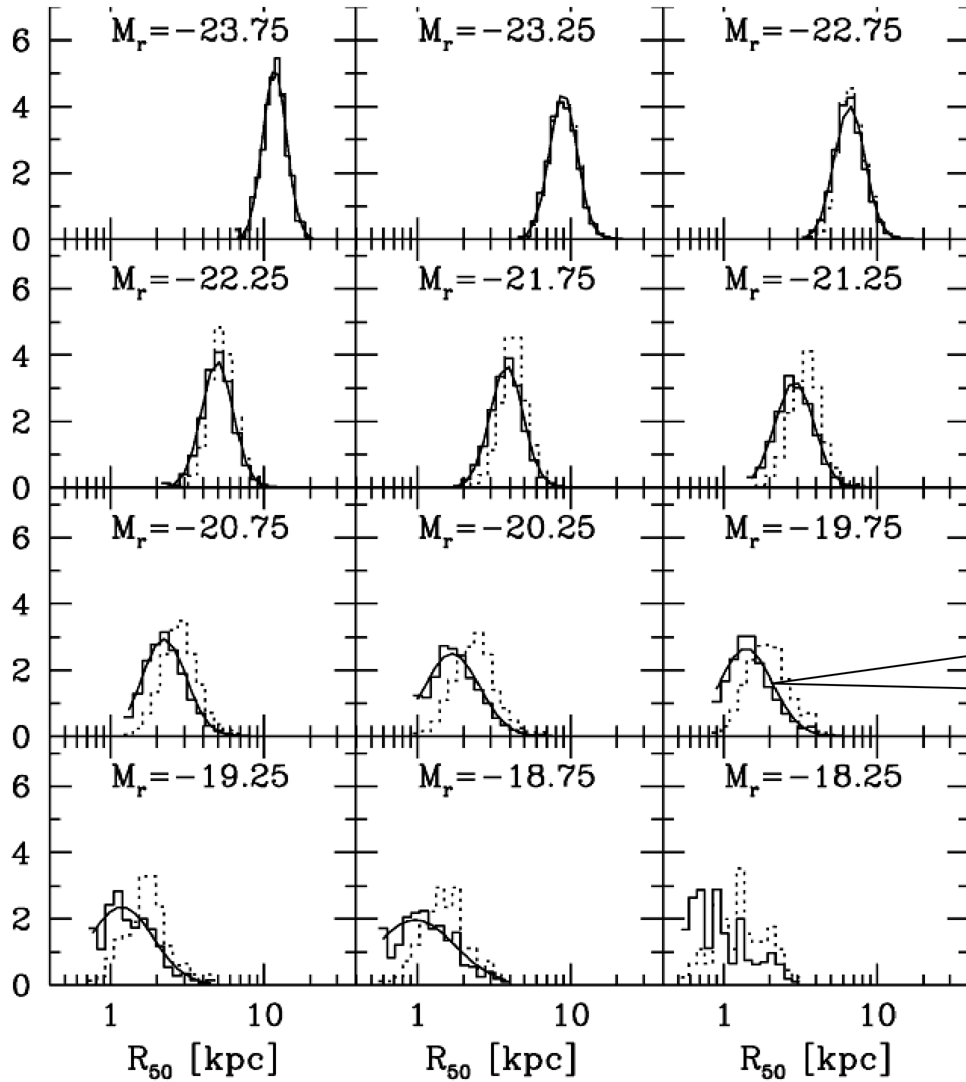
- Especially when amount of data is large in modern surveys

- First, at given bin of x, what is the distribution of y after correction for selection bias?
  - Is y Gaussian distributed? What is the scatter compared with its measurement error?

- Then what is the PDF(y|x) changes as function of x
  - Is this relation linear or non-linear?

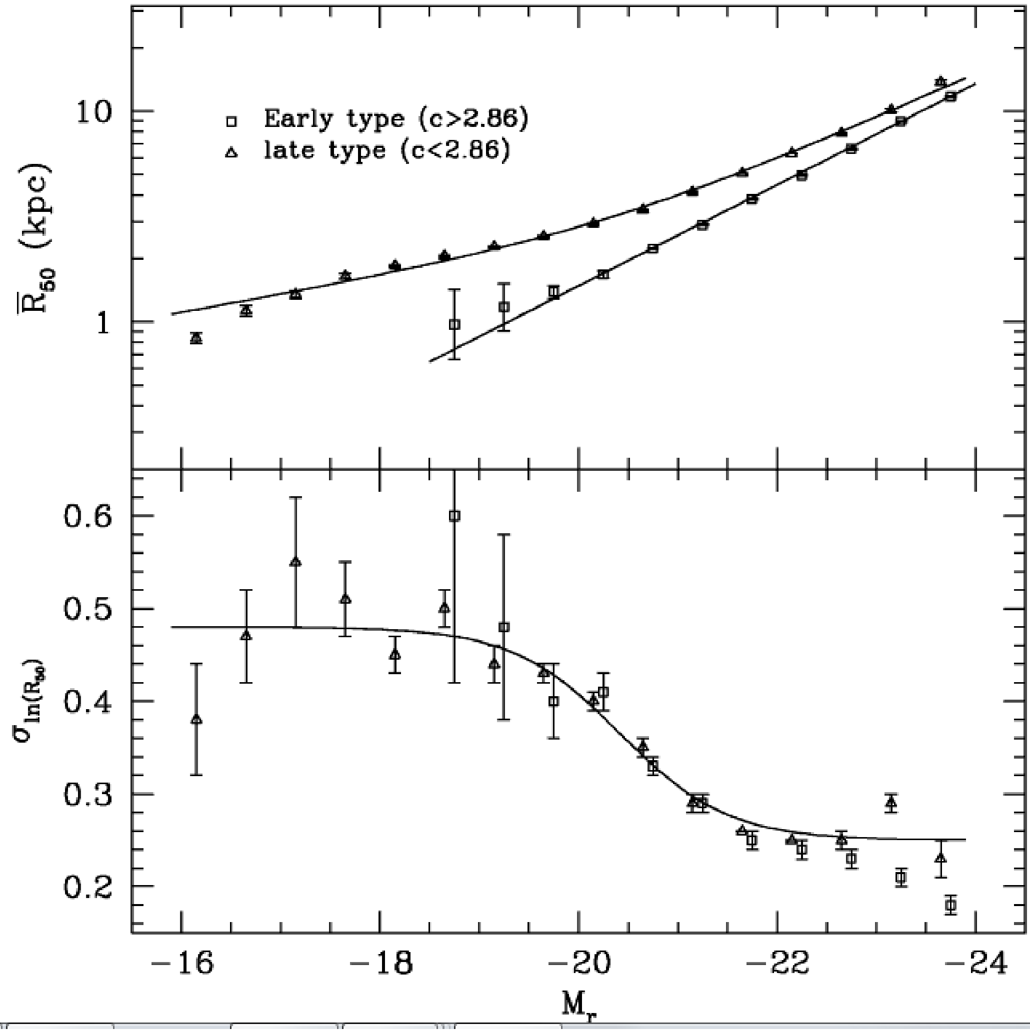
- Build the likelihood function and fit the model parameters

L – R relation of galaxies (Shen et al. 2003)



We find, after correction for selection effect, at given Mr, Log *R* is intrinsically Gaussian distributed.

Data is biased here



We plot P(R|M) as function of M.

Intrinsic scatter is not a constant

I.5: correlations between parameters

Is the correlation between A and B real or

because A and B are both correlated with C?

Partial correlation

- X correlated with Z, Y correlated with Z, whether X correlated with Y
  - Distance dependent parameters, e.g. L_R VS L_X
- Idea: calculate the correlation between the residuals
  - assumes linear relationship. $r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$.
- More generalized: multiple regression

Control sample

- We see different b/a values between AGNs and normal spirals. What does it mean? (Shen et al. 2010)
  - b/a is function of stellar mass, size etc.
  - AGNs biased to high stellar mass sample
- We build a control sample of galaxies, which have the same stellar mass, size, concentration, color distributions as AGNs
  - We then compare the b/a of AGNs with control sample

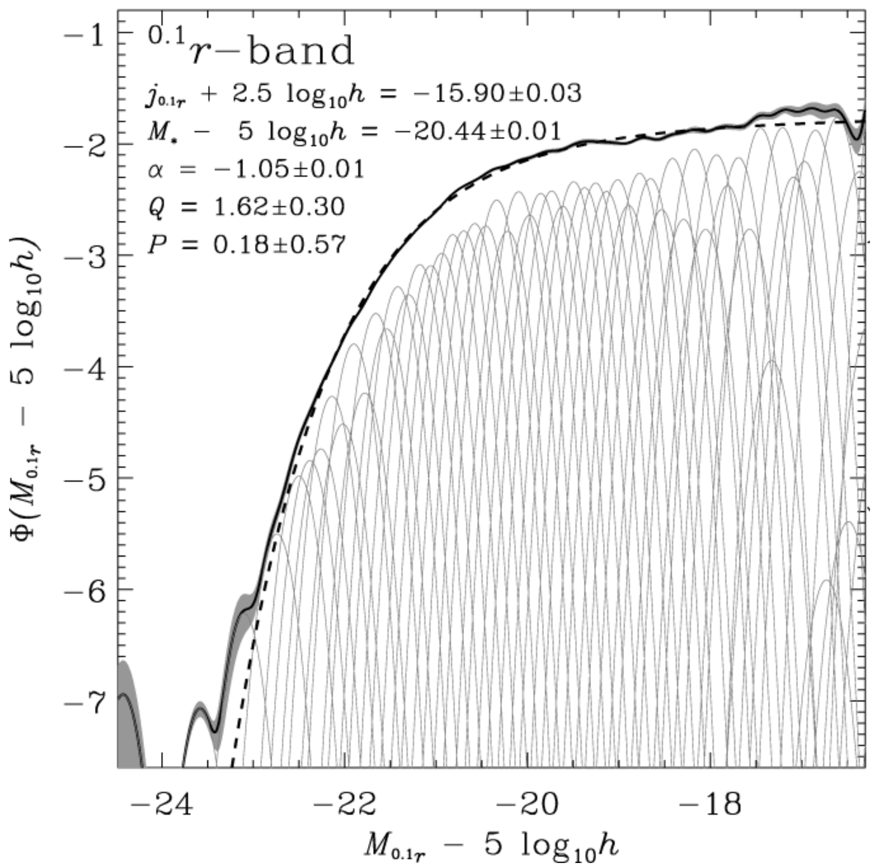
II. Luminosity function of galaxies

LF of galaxies

- The basic statistical properties of galaxies in any galaxy survey
- Schechter function
  - Characteristic luminosity M_*
  - Faint end slope α

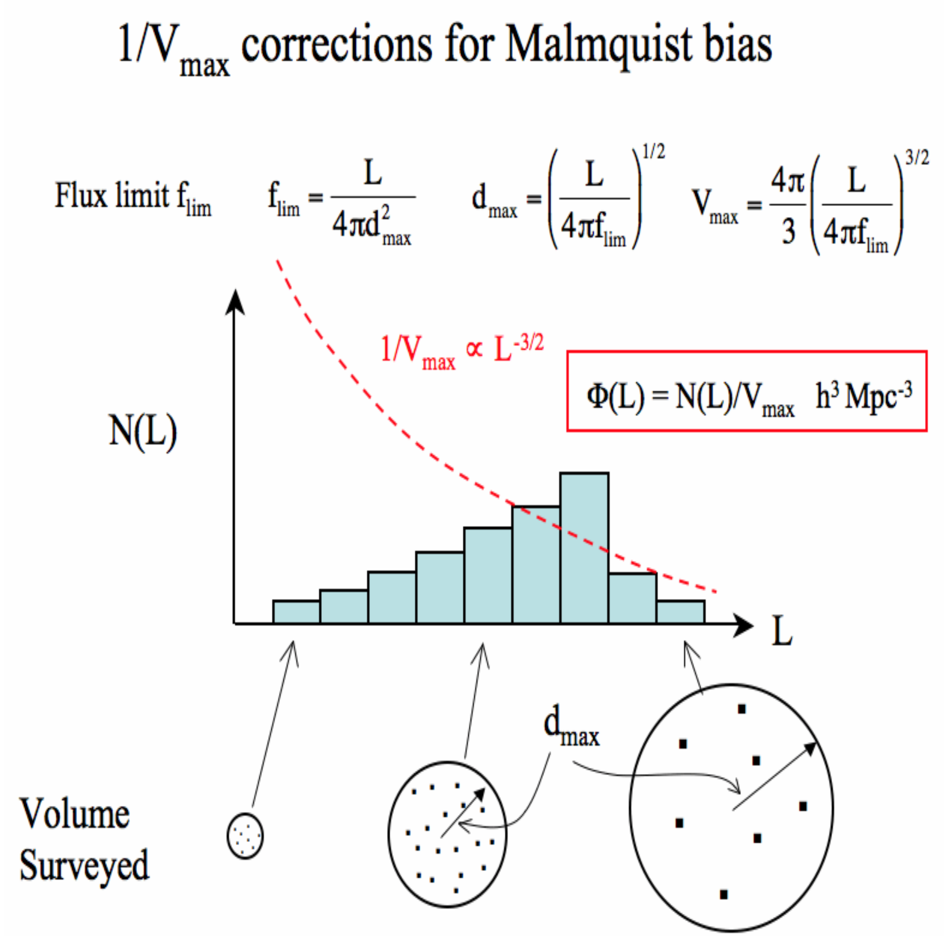
$$\phi(L)dL = \phi^*\left(\frac{L}{L^*}\right)^{\alpha} exp\left(-\frac{L}{L^*}\right)\frac{dL}{L^*}$$

Blanton et al. (2003) (astro-ph/0210215)



Traditional Vmax estimation of LF(Felton 1977)

- Vmax: maximum volume of a galaxy with certain absolute luminosity can be observed in the flux limited sample
  - For flux limit complete sample: <V/Vmax>=0.5
- Advantage: no assumption of the LF shape
- Shortcoming: based on the assumption that galaxy distribution is homogenous



Maximum likelihood estimation

- The probability of a galaxy in the sample

$$p_i = \left(\frac{\Phi(L_i)}{\int_{L_{min}(d_i)}^{\infty} \Phi(L)\, dL} \right)$$

$$\phi(L)dL = \phi^* \left(\frac{L}{L^*}\right)^{\alpha} exp\left(-\frac{L}{L^*}\right)\frac{dL}{L^*}$$

  - $L_{min}(d_i)$, the minimum luminosity above the flux limit.
    - Selection effect
- The likelihood function $\boldsymbol{P} = \prod_i p_i$
- Maximize L as function of M_*, α
  - How to maximize?
    - Analytical: exercise on a Gaussian distribution.
    - numerical calculations in parameter space
  - No direct constraint on ϕ_*

$$\frac{\partial lnP}{\partial \alpha} = 0$$

$$\frac{\partial lnP}{\partial L^*} = 0$$

Step-Wiese Maximum Likelihood method (Efstathiou et al. 1988)

- LF is function of N steps
  - Avoid to use Schechter function as a prior

$$\phi(L) = \phi_k, \quad L \in (L_k - \Delta L/2, L_k + \Delta L/2), \quad k = 1, \ldots, N$$

The likelihood, as in the previous method, then is:

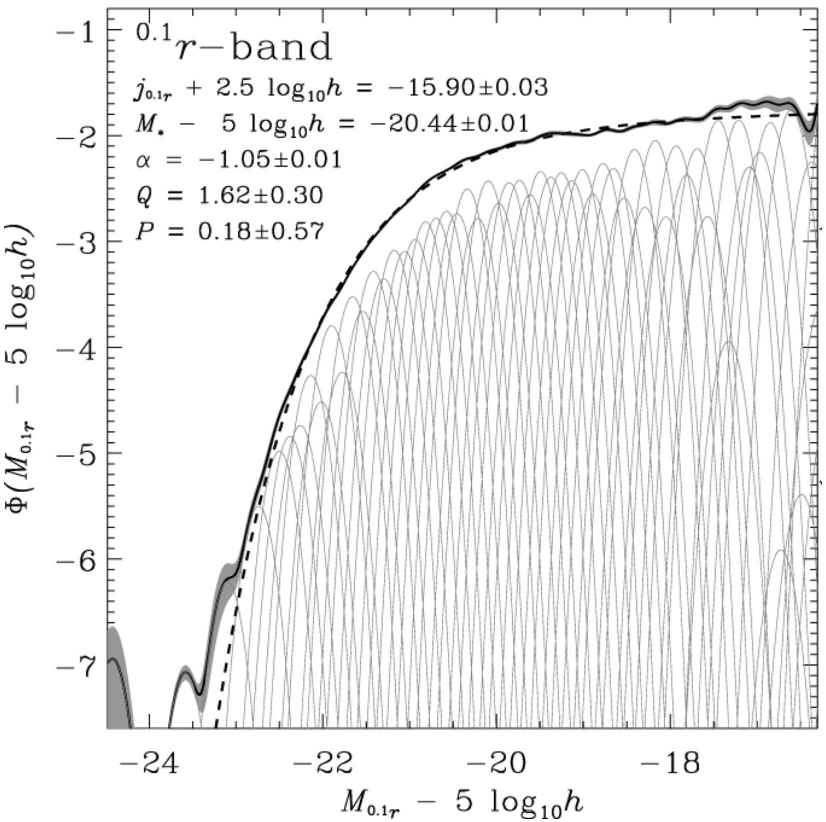
$$\ln L = \sum_{i=1}^{N} W(L_i - L_k) \ln \phi_k - \sum_{i=1}^{N} \ln\{\sum_{j=1}^{N} \phi_j \Delta L H[L_j - L_{min}(z_i)]\} + C$$

LF estimator of SDSS (Blanton et al. 2003)

$$\Phi(M,z) = \bar{n}10^{0.4(z-z_0)P} \sum_k \Phi_k \frac{1}{\sqrt{2\pi\sigma_M^2}} \times \exp\left\{-\frac{1}{2}\frac{[M - M_k + (z-z_0)Q]^2}{\sigma_M^2}\right\}$$

- Using *n* Gaussian instead of steps
- Considering luminosity evolution (Q)

Blanton et al. (2003) (astro-ph/0210215)



Notes on LF estimation

- Sample completeness is most important
  - Low surface brightness galaxies are always the topic
- Should consider cosmic variance in high redshift survey
- With modern data, conditional LFs are discussed more and more
  - Morphology, color, environment etc.

III. Stellar synthesis model

Stellar population synthesis model

- What can we say about the star formation history of galaxy from photometric colors or spectroscopy?

- Key elements
  - Physics of stellar evolution [function (M,z)] is classic
    - either empirical or theoretical grounds
  - Single-age, single-metalicity population (SSP)
    - Linked to stellar isochrones with a statistical parameter IMF (Initial mass function)
  - Stellar populations + other ingredients (e.g. dust attenuation, kinematics, HII regions, AGN) → observed galaxy properties

Composite stellar populations (CSPs)

- Galaxy composited by several SSPs

$$F_\lambda(t,Z) = \int_0^t \Psi(t-t')S_\lambda(t',Z)\,e^{-\tau_\lambda(t')}\,\mathrm{d}t'$$

A galaxy made of two populations

$$F_\lambda \sim \Phi(M_1)\Psi(t_1)S_\lambda(M_1,t_1) + \Phi(M_2)\Psi(t_2)S_\lambda(M_2,t_2)$$

linear regression of the stellar populations

- Output: color, spectra indices etc. y_i
- Components: M SSPs of different age X_k
- Coefficient: a_k
- Linear regression 'lfit' in Numeric recipes

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2 \longrightarrow 0 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=1}^{M} a_j X_j(x_i) \right] X_k(x_i) \qquad k = 1, \ldots, M$$

$$\sum_{j=1}^{M} \alpha_{kj} a_j = \beta_k \qquad \alpha_{kj} = \sum_{i=1}^{N} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \qquad \beta_k = \sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2}$$

- However, a_i can not be negative
  - Non-negative linear regression: also applied on image analysis
  - IDL (Fortran) code: NNLS Lawson and Hanson (1983)
- Improved version: BVLS
  - solves linear least-squares problems with upper and lower bounds on the variables

Choose your evidence

- If we have a spectrum of galaxy, what do we use as the constraints of stellar populations?
  - More features are better, more information, more constraints
  - However, evidence may tell more than theory
    - Emission lines should be removed
    - UV continuum should be careful, dust
    - SSP library maybe limited at certain wavelength
      - E.g. BC03 VS BC07
- Choose the proper evidence to quantify specific question
  - So science is not just statistics

Reference: Comparing six evolutionary population synthesis models through spectral synthesis on galaxies (Chen et al. 2010)

Stellar synthesis: evolved stellar population

- We have prior information about star formation history (SFH) of galaxies
  - ☹ an old/metal rich stellar population + a young/metal poor stellar population
  - We know the cosmic star formation history (Madu plot)
- we may parameterize the SFH of galaxies in simple way, e.g. SFH=$e^{-t/\tau}$ (e.g. BC03)
  - for limited evidence (data), e.g. only color
  - Assumption may be too simplistic, but physics is there

Reference: stellar mass of SDSS galaxies, Kauffmann et al. 2003

SFH: Bayesian approach

- P(SFH|Spec)=P(SFH)*P(Spec|SFH)/P(Spec)
  - Evidence: P(Spec)=1
  - P(Spec/SFH): estimated from χ^2
  - Prior: P(SFH) ?
- Build *K* SFH libraries
  - $\Sigma_{i=1,K} P(SFH_i)=1$
  - $P(SFH_i)/P(SFH_j)=P(Spec|SFH_i)/P(Spec|SFH_j)$
- Assumptions:
  - equal prior for each library
  - Library cover all possibilities

Numeric Simulation, Semi-analytic model, phenomenological model

- Numeric simulation: include as more known physics as possible
  - But can not include all, anything new?
  - SFH: N-body (dark matter) + SPH(gas)
- Semi-analytic model: based on some results from simulation, parameterize some unknown physical process, e.g. star formation
  - SFH: halo merge history (from simulation)+ parameterized star formation law
- Parameterize the complicate physical process e.g. SFH=$e^{-t/\tau}$

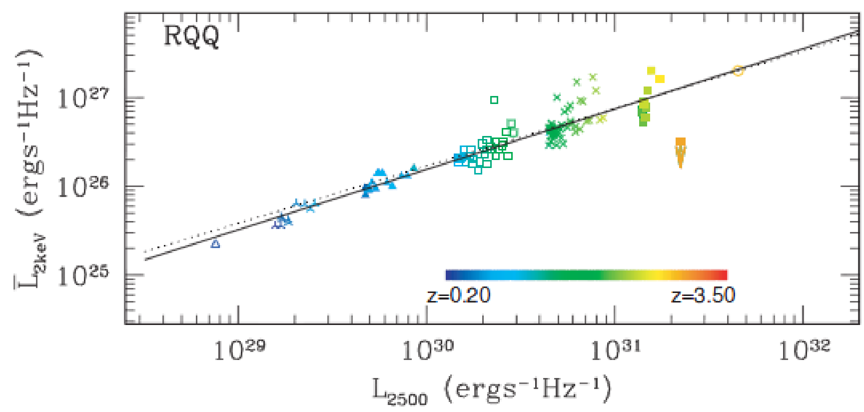
IV: stacking technique

- Only upper limits for very faint source
  - needs deeper exposure
- Upper limit includes information
- Stacking: sources supposed to share similar properties, stacking then is equivalent to increase the exposure time
  - Space → time
  - get average properties
    - Signal may be dominated by few bright sources

Mean VS median

- Mean L_{2KeV} at given L_{2500} in stacking
- Median L_{2KeV} at given L_{2500} in individual linear fitting
  - Fitting in Log L_{2KeV}-Log L_{2500} space
- Scatter of Log L_{2KeV} is ~0.4
  - mean and median difference is a factor of 1.7
- Answer maybe the quasar variability
  - Log-normal

Excellent agreement between stacks and individual detection here is misleading



Solid: data from stacks of QSO.
Dotted: data from individual detection.
Shen et al. 2006

V: Extreme value statistics

- Extreme value populations are easily observed
  - e.g. the brightest group/cluster galaxies, the brightest star of a star cluster
    - Order statistics of the early-type galaxy luminosity function (Dobos & Csabai 2012)
- What can a extreme value tell us ?
  - How unusual are the Shapley Supercluster and the Sloan Great Wall (Sheth & Diaferio 2011)
  - Quantifying the rareness of extreme galaxy clusters (Hotchkiss 2011)
  - An application of extreme value statistics to the most massive galaxy clusters at low and high redshift (Waizmann, Ettori, & Moscardini 2012)
  - Temperature maximum in CMB (coles 1988)

Extreme value statistics

- Three types of extreme value distribution, Depends on the tail shape (Fisher–Tippett–Gnedenko theorem)
  - Weibull(no tail)
    - Lowest temperature
  - Fréchet(flat tail)
    - Money of richest people
  - Gumbel (exponential tail)
    - Height of people
  - Requires sample size N>>1
- Brightest group/cluster galaxy
  - Gumbel distribution?

Extreme value statistics/Order statistics (EVS/OS Dobos & Csabai 2011)

- Cumulative distribution of distribution function $f(x)$

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u.$$

- probability of a number $x < X$

$$P(x < X) = F(X).$$

- N independently drawn numbers $\{x_1, x_2, \ldots, x_N\}$, the probability of $\max\{x_i\} = X_m$

$$P_{\mathrm{m}}(X_{\mathrm{m}}) = P(x_i < X_{\mathrm{m}}) = P^N(x < X_{\mathrm{m}}) = F^N(X_{\mathrm{m}}).$$

- the probability density function of the maximum of a sample of size N

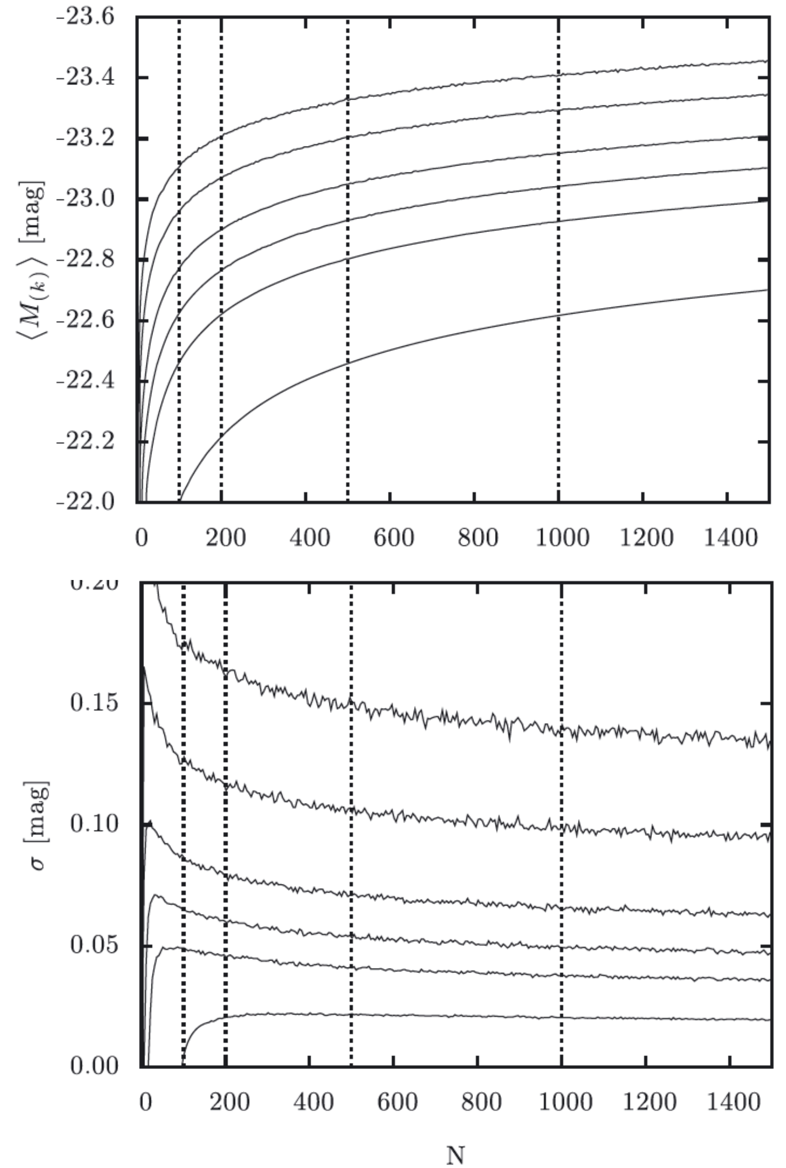
$$p_{\mathrm{m}}(X_{\mathrm{m}}, N) = N F^{N-1}(X_{\mathrm{m}}) f(x),$$

- The probability distribution of the kth largest value

$$p_{(k)}(X_{(k)}, N) = \frac{N!}{(k-1)!(N-k)!}[1 - F(X_{(k)})]^{k-1} F^{N-k}(X_{(k)})\, f(X_{(k)}).$$

EVS/OS: basic conclusions

- The mean extreme values of a lager sample is larger
  - Height of Chinese basket-ball team player is taller than Japanese
  - Brightest galaxies of rich clusters is more luminous than poor groups
- The scatter of the extreme values of a lager sample is smaller
  - BCGs have small scatter
  - The scatter of the higher order members is even smaller



Dobos & Csabai 2011

Other advanced topics not listed

- Principle component analysis (PCA)
  - In spectrum analysis
- Fourier transform
  - Image analysis
  - Time series
- Monte-Carlo Markov chain
  - Find the best model parameters in multi-dimensional space
- Data mining
  - Virtual Observatory
- etc…

Final thoughts

- Use proper model
  - Depend on your question.
  - Question is the first step of your science
- Use proper way to do the statistics
  - Need to know the principle, may need not know the detail.
- Use proper evidence
  - Model explains everything is wrong
  - Depend on your knowledge and experience
- Data mining