BASIC STATISTICS III

CHI SQUARED AND LIKELIHOOD

Individual values of Chi squared – comparison of observed residuals from a model to expected residuals (uncertainties)

$$\chi^2 = \sum_i \frac{(O_i \quad E_i)^2}{\sigma_i^2}$$

These individual values come from an underlying Chi squared probability distribution.

The likelihood of a model given a data set is *proportional to* $e^{-\chi^2/2}$ because we assume the residuals follow normal distributions.

FITTING

- traditional maximum likelihood:
  “best fit” models/parameters
- the right fit for the question
- the Bayesian approach:
  probability *distributions* for models/parameters

TRADITIONAL MAXIMUM LIKELIHOOD

- likelihood proportional to $e^{-\chi^2/2}$
- min χ^2 ➔ max likelihood
- "maximum likelihood estimators" (MLEs) of parameters α_i of a model are usually found by $\frac{\partial L}{\partial \alpha_i} = 0$ or equivalently $\frac{\partial \ln(L)}{\partial \alpha_i} = 0$
- if residuals are Gaussian, called "ordinary least-squares" (OLS) fitting – minimizes rms deviations

TRADITIONAL MAXIMUM LIKELIHOOD

- Example: model $y = \alpha X + \beta$ with equal Gaussian errors σ
- $\chi^2 = \sum_i \frac{(Y_i - (\alpha X_i + \beta))^2}{\sigma^2}$ ➔ max likelihood
- $\frac{\partial \ln(L)}{\partial \alpha} = 0$ ➔ $\frac{\partial \ln\left(e^{-\frac{\chi^2}{2}}\right)}{\partial \alpha} = 0$ ➔ $\sum_i \frac{(Y_i - (\alpha X_i + \beta))X_i}{\sigma^2} = 0$
- $\frac{\partial \ln(L)}{\partial \beta} = 0$ ➔ $\frac{\partial \ln\left(e^{-\frac{\chi^2}{2}}\right)}{\partial \beta} = 0$ ➔ $\sum_i \frac{(Y_i - (\alpha X_i + \beta))}{\sigma^2} = 0$
- two eqns, two unknowns – solve to get result in tutorial:

  $\alpha = \frac{\bar{X}\bar{Y} - \overline{XY}}{(\bar{X})^2 - \overline{X^2}}$ and $\beta = \bar{Y} \quad \bar{X}\alpha$

  (so for this simple case, no numerical χ^2 minimization is needed; harder for more parameters or different σ_i)

TRADITIONAL MAXIMUM LIKELIHOOD

- uncertainties on MLEs estimated by 1/E(-H) = inverse of expectation of negative "Hessian matrix"
- Hessian matrix example: $y = \alpha X + \beta$

$$\text{Hessian}(\alpha, \beta) = \begin{bmatrix} \frac{\partial^2}{\partial\alpha^2}\log L(\alpha, \beta) & \frac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha, \beta) \\ \frac{\partial^2}{\partial\alpha\partial\beta}\log L(\alpha, \beta) & \frac{\partial^2}{\partial\beta^2}\log L(\alpha, \beta) \end{bmatrix}$$

note covariance terms!

- complicated to compute Hessians, often done numerically
- fully worked Hessian for least squares case at http://mathworld.wolfram.com/LeastSquaresFitting.html - note errors on parameters generally decrease as $\frac{1}{\sqrt{N}}$

TRADITIONAL MAXIMUM LIKELIHOOD

Even simple line fitting gets complicated really quickly…

- Isobe+ (1990): OLS(y|x), OLS(x|y), bisector fit
- Beers+ (1990): biweight for robust (outlier-resistant) fitting
- A. Trotters' 2011 UNC PhD thesis: "I present a new, broadly applicable statistical technique… for fitting model distributions to data in two dimensions, where the data have intrinsic uncertainties in both dimensions, and extrinsic scatter in both dimensions that is greater than can be accounted for by the intrinsic uncertainties alone."

THE RIGHT FIT FOR THE QUESTION

- Hogg, Bovy, and Lang (2010) model fitting review (http://arxiv.org/pdf/1008.4686v1.pdf) claims you must choose OLS(y|x) or OLS(x|y) based on comparing σ_x and σ_y
- fair enough if you're after the underlying relationship between x & y, but what if that's not your question? (see earlier Feigelson & Babu (1992) fitting review)

EXAMPLE: For the Tully-Fisher Relation,

1) which fit is best if you want a relation useful to *predict* L, and in this case how should you trim the data?
2) which fit is best if you want to study the residual dependence on third parameters at fixed L, and again how should you trim the data?
3) which fit is best if you want to study the TFR itself, assuming V has the most scatter, and there is selection bias on L?

SELECTION EFFECTS, BIAS CORRECTION, AND SURVIVAL ANALYSIS

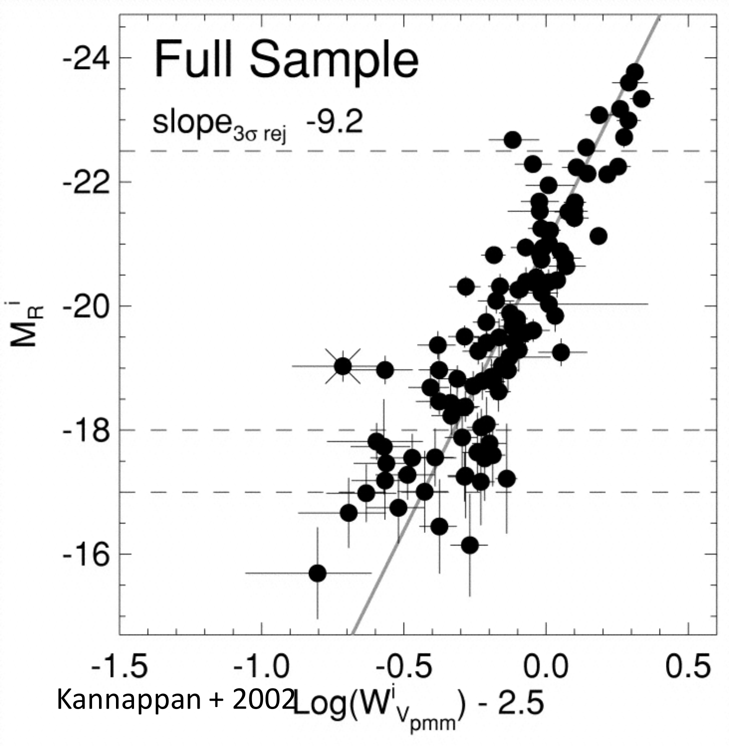
Censoring: non-detections that can be displayed as upper or lower limits (e.g. HI gas non-detections)

Truncation: non-detections that cannot be displayed; numbers of data points and/or limiting values unknown (faint end of Tully-Fisher relation)

Q: How can we fit censored/truncated data?

A1: Use initial data model (with scatter) to replace limits/missing data with Monte Carlo data; iterate fit-MC-fit-MC to convergence ("survival analysis")

A2: Construct data model with scatter and selection limit and iteratively "bias correct"; use simulated data as models (Approximate Bayesian Computation)



Kannappan + 2002

THE BAYESIAN APPROACH: ~~MAXIMUM~~ LIKELIHOOD DISTRIBUTIONS

- construct space of models with expected ranges of parameters
- prob(model|data) ∝ prob(data|model) × prob(model)
  or Posterior ∝ Likelihood × Prior
- for flat priors on model params, posterior probability of each model is proportional to its likelihood, $e^{-\chi^2/2}$ (math matches frequentist, but Bayesian would keep full posterior and not make a "point estimate")
- integrate over "nuisance parameters" ("marginalize" over them) to get probability distribution for one parameter

THE BAYESIAN APPROACH: ~~MAXIMUM~~ LIKELIHOOD DISTRIBUTIONS

More uses of marginalization:

- Compare whole classes of models A and B (e.g. straight line or curved?) by integrating over parameters:

$$\text{“Bayes Factor”} = \frac{\int_\alpha p_A(X_i|\alpha,A)p(\alpha|A)}{\int_\alpha p_B(X_i|\alpha,B)p(\alpha|B)}$$

- Get probability distributions for model quantities other than model parameters, e.g., galaxy stellar masses from SPS fits to galaxy SEDs (stellar mass is not an input to the model, rather it is a scale factor determined in the fit)

THE BAYESIAN APPROACH: CHOICE OF PRIORS

- Priors can be “uninformative” in different ways – e.g. uniform in linear parameter, “scale-free” (uniform in log parameter), uniform in a transformed parameter
- Given lots of data and/or small error bars, posterior results should insensitive to the choice of prior within the spread in the posterior
- Given crappy data or few data points, should test sensitivity of posterior to reasonable priors – and the full Bayesian posterior is crucially more informative in such a case than a frequentist point estimate
- Example: paramfit2_nrun.py