Chi-square, K-S, and bootstrap: Fitting astrophysical models to data

G. Jogesh Babu

Penn State University
https://science.psu.edu/stat/people/gjb6

http://astrostatistics.psu.edu

Eberly College of Science

PennState

Center for Astrostatistics

Model Fitting

- Non-linear regression
- Density (shape) estimation
- Parametric modeling
- Goodness of fit

Model Fitting in Astrophysics

- Interpreting spectrum of an accreting black hole such as a quasar.
- Interpreting radial velocity variations of a large sample of solar-like stars.
- Interpreting spatial fluctuations in cosmic microwave background radiation.

- Are there any interesting correlations among the properties of objects in any given class (e.g. the Fundamental Plane of elliptical galaxies), and what are the optimal analytical expressions of such correlations?

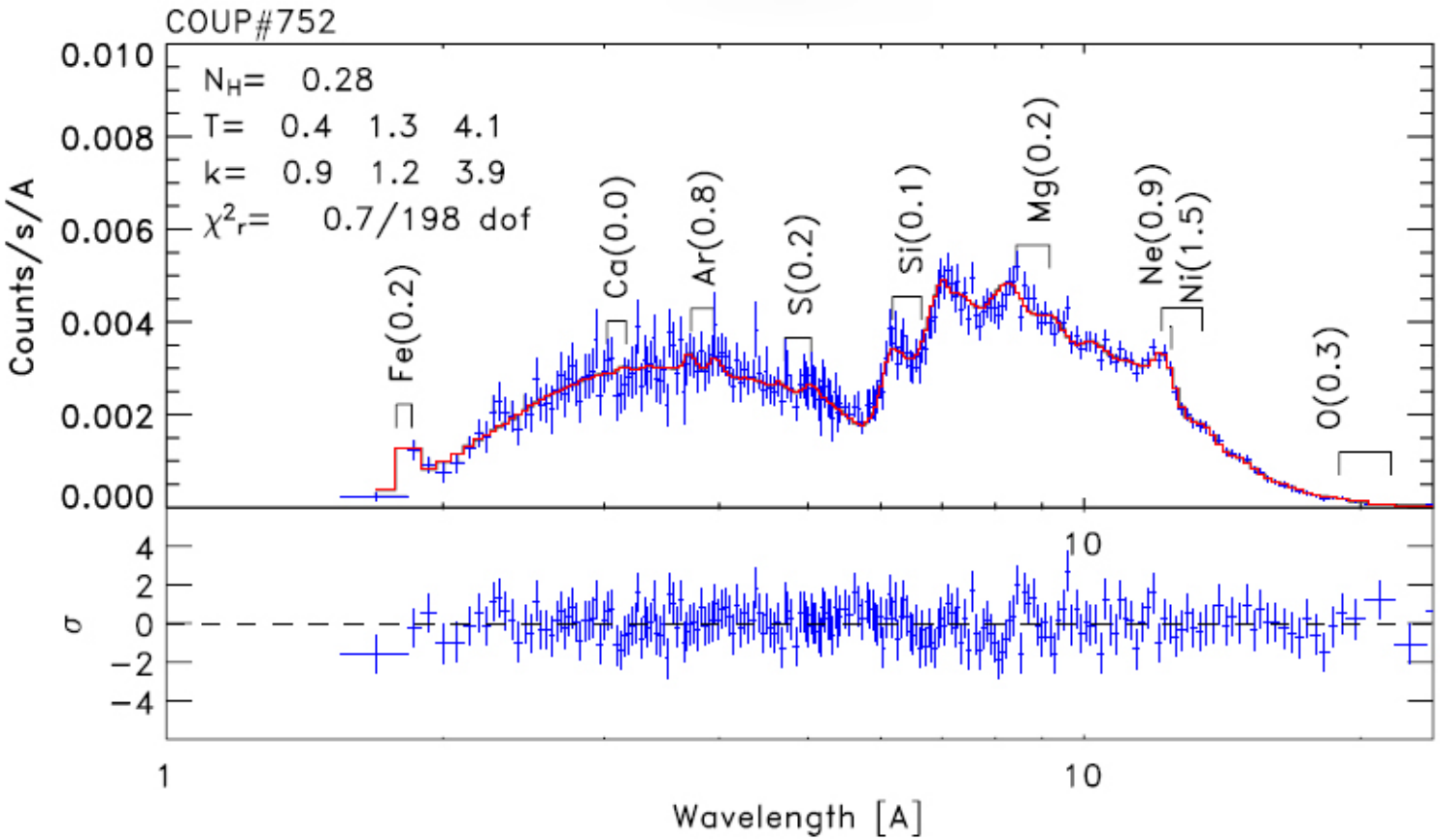
These issues arise when data are used to repudiate or support astrophysical theories but the underlying processes generating the data are not confidently known.

A good model should be

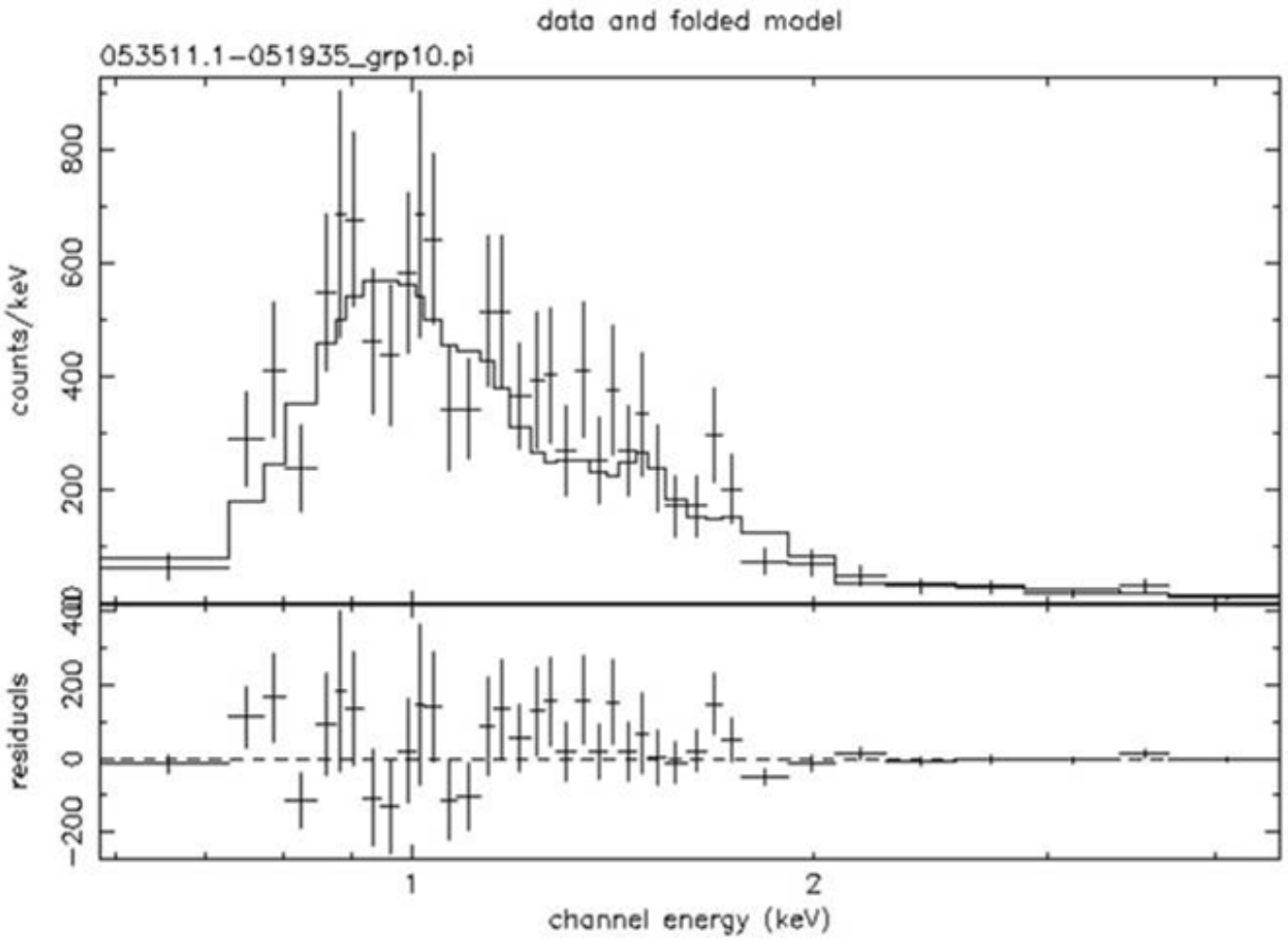
- Parsimonious (model simplicity)
- Conform fitted model to the data (goodness of fit)
- Easily generalizable.
- Not *under-fit* that excludes key variables or effects
- Not *over-fit* that is unnecessarily complex by including extraneous explanatory variables or effects.
- Under-fitting induces bias and over-fitting induces high variability.

A good model should balance the competing objectives of conformity to the data and parsimony.

Successful model for high signal-to-noise X-ray spectrum



A bright source from Chandra Orion Ultradeep Project
Complicated thermal model with several temperatures
and element abundances (17 parameters)



COUP source # 410 in Orion Nebula with 468 photons
Fitting binned data using χ^2
Model with three parameters (θ):
plasma temperature; line of sight absorption; normalization

A plausible emission mechanism

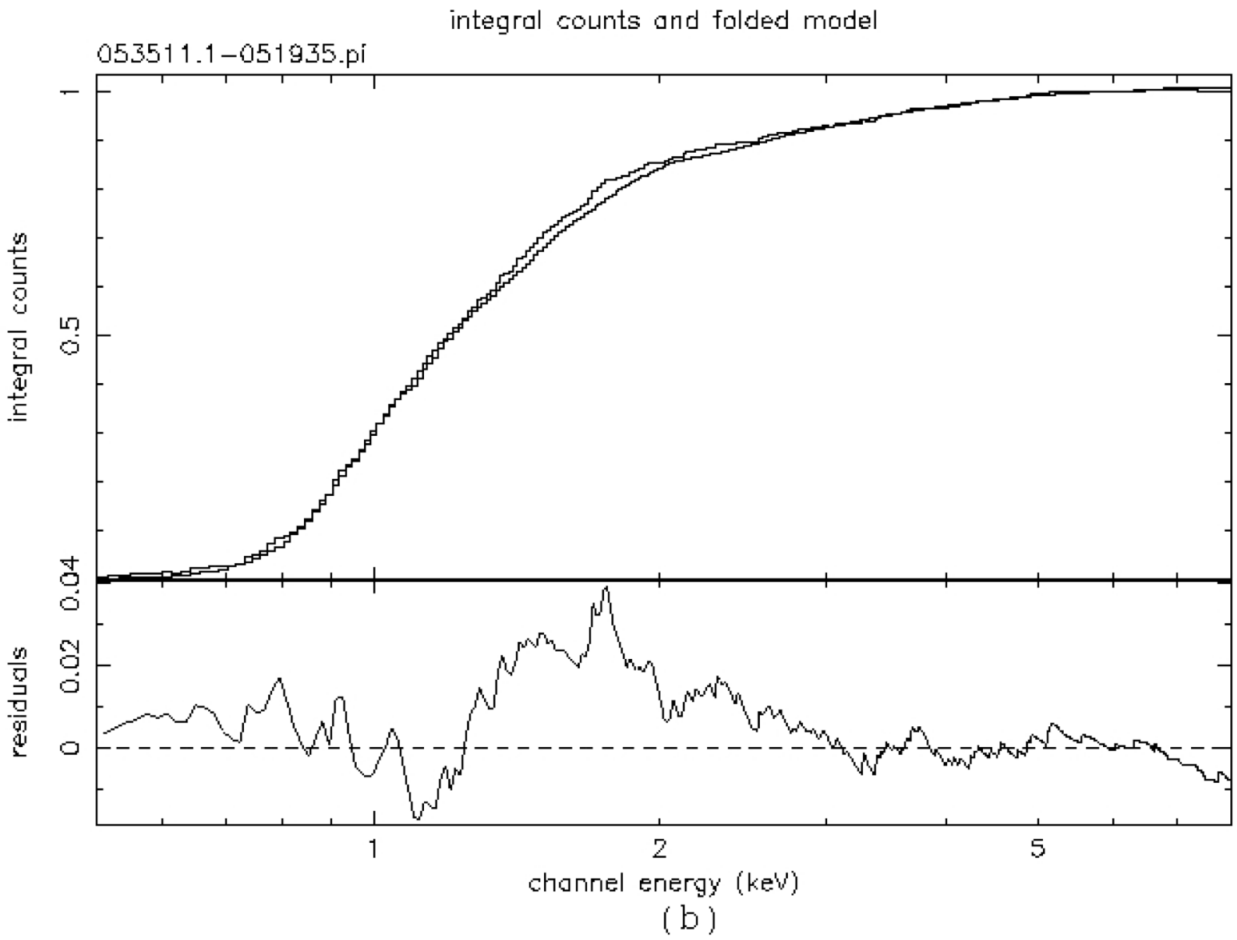
- The astrophysical model has been convolved with complicated functions representing the sensitivity of the telescope and detector.
- The model is fitted by minimizing sum of squares ('minimum chi-square') with an iterative procedure. (Bevington 1969)

$$\hat{\theta} = \arg\min_{\theta} \chi^2(\theta) = \arg\min_{\theta} \sum_{i=1}^{N} \left(\frac{y_i - M_i(\theta)}{\sigma_i} \right)^2 .$$

It is parameter estimation by *weighted least squares*.

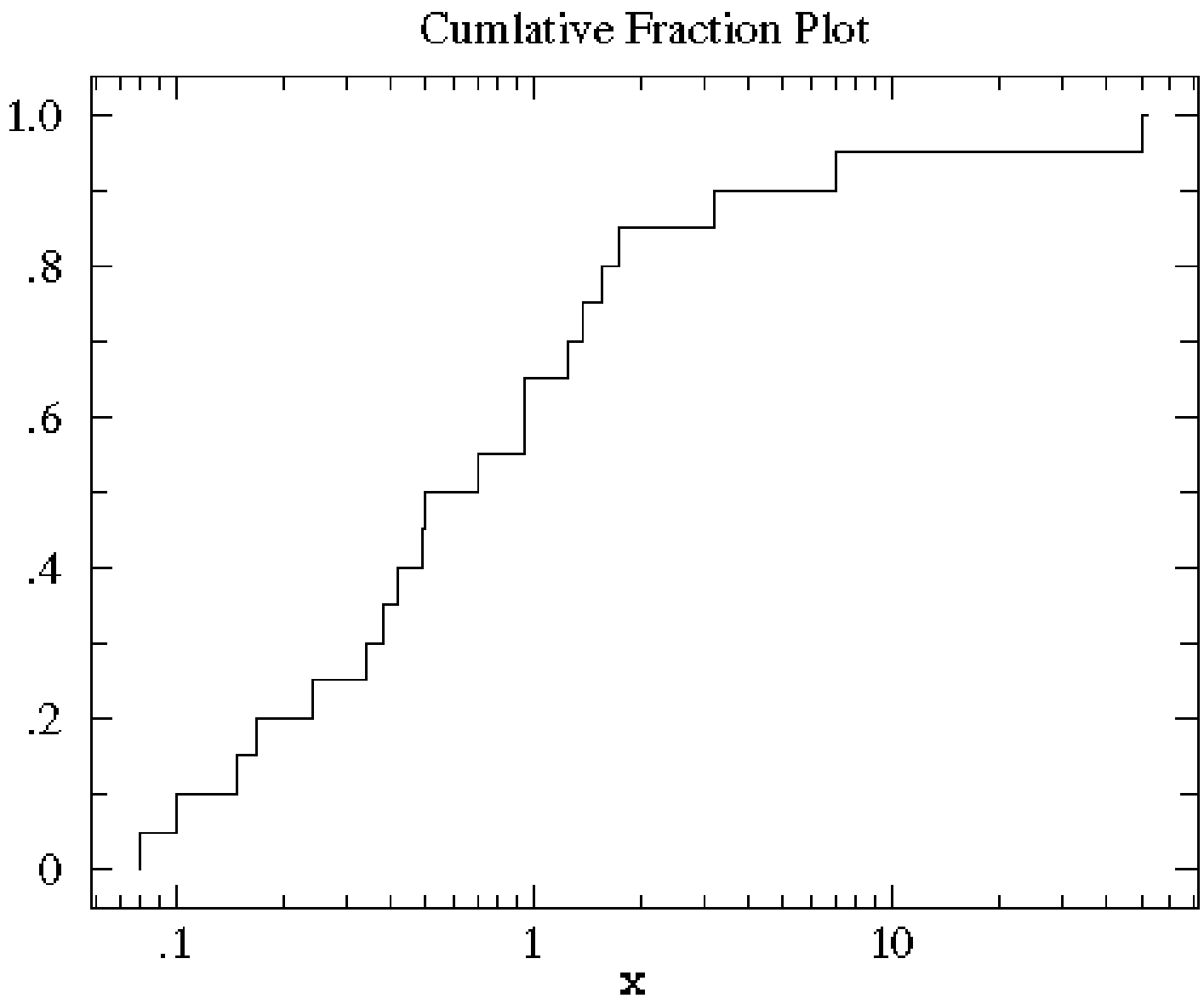
Limitations to *weighted least squares*

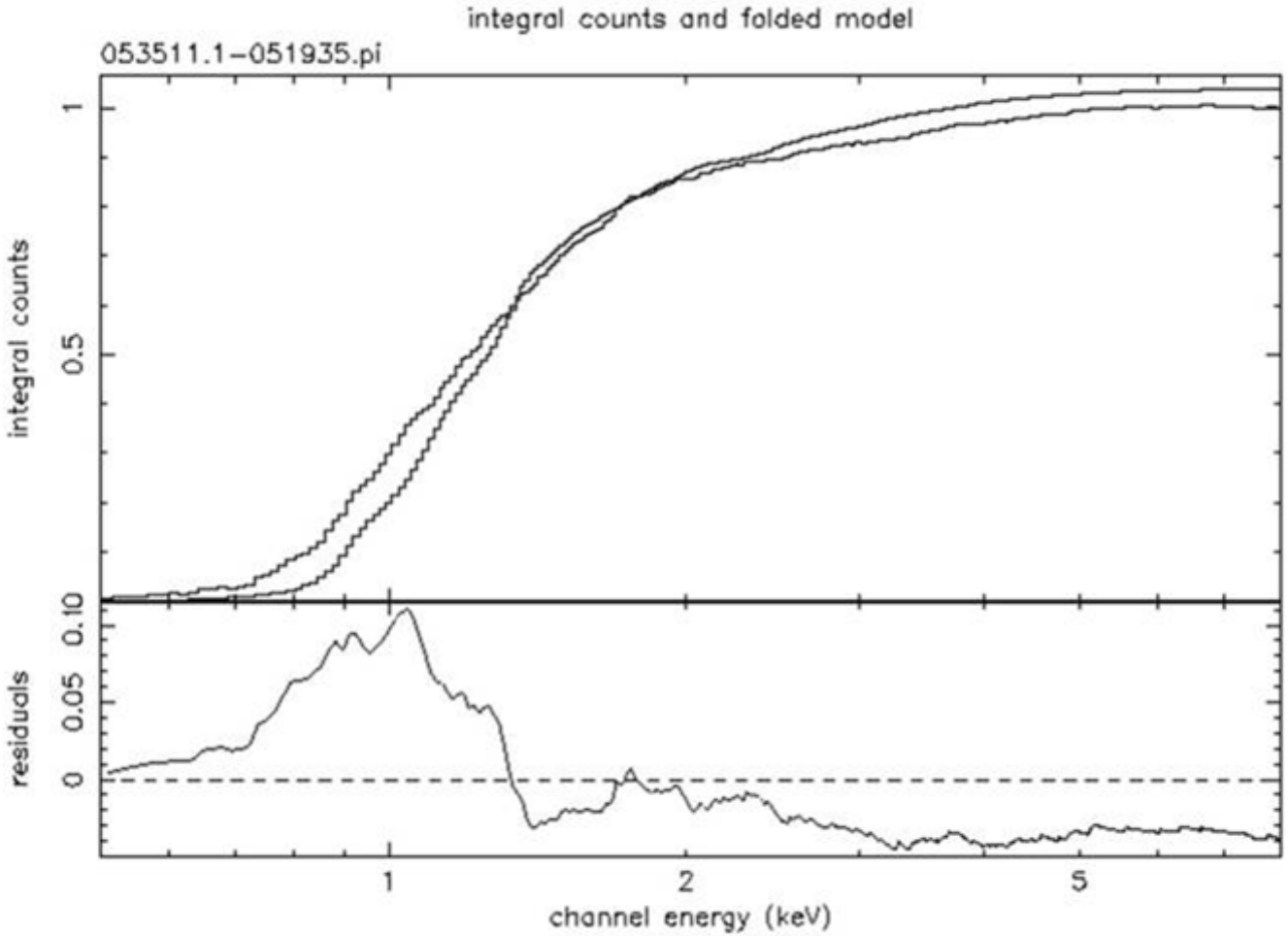
- Fails when bins have too few data points.
- Binning is arbitrary. Binning involves loss of information.
- Data points should be independent. Failure of independence assumption is common in astronomical data due to effects of the instrumental setup.
- Does not provide clear procedures for adjudicating between models with different numbers of parameters.



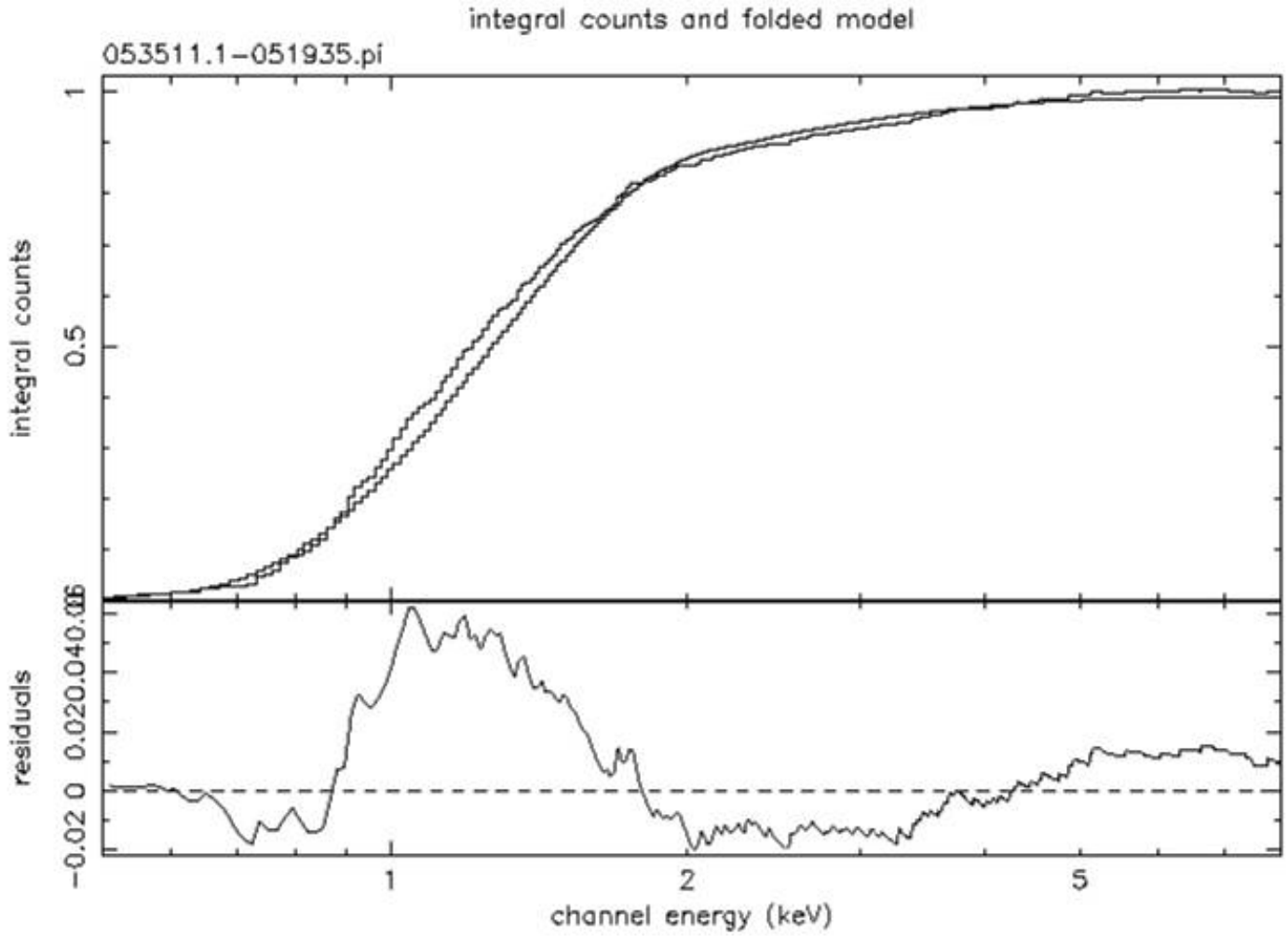
Thermal model with absorption
Fitting to unbinned data
Maximum likelihood (C-statistic)

Empirical Distribution Function





Fitting to unbinned EDF
Correct model family, incorrect parameter value
Thermal model with absorption set at $A_V \sim 10$ mag
What is the 99%confidence interval for A_V?



Misspecified model family!
Power law model with absorption set at $A_V \sim 1$ mag
Can the power law model be excluded with 99% confidence

Outline

Statistics based on EDF

Kolmogrov-Smirnov: $D_{n,F} = \sup_x |F_n(x) - F(x)|$.

The sampling distribution of K-S statistic is given by

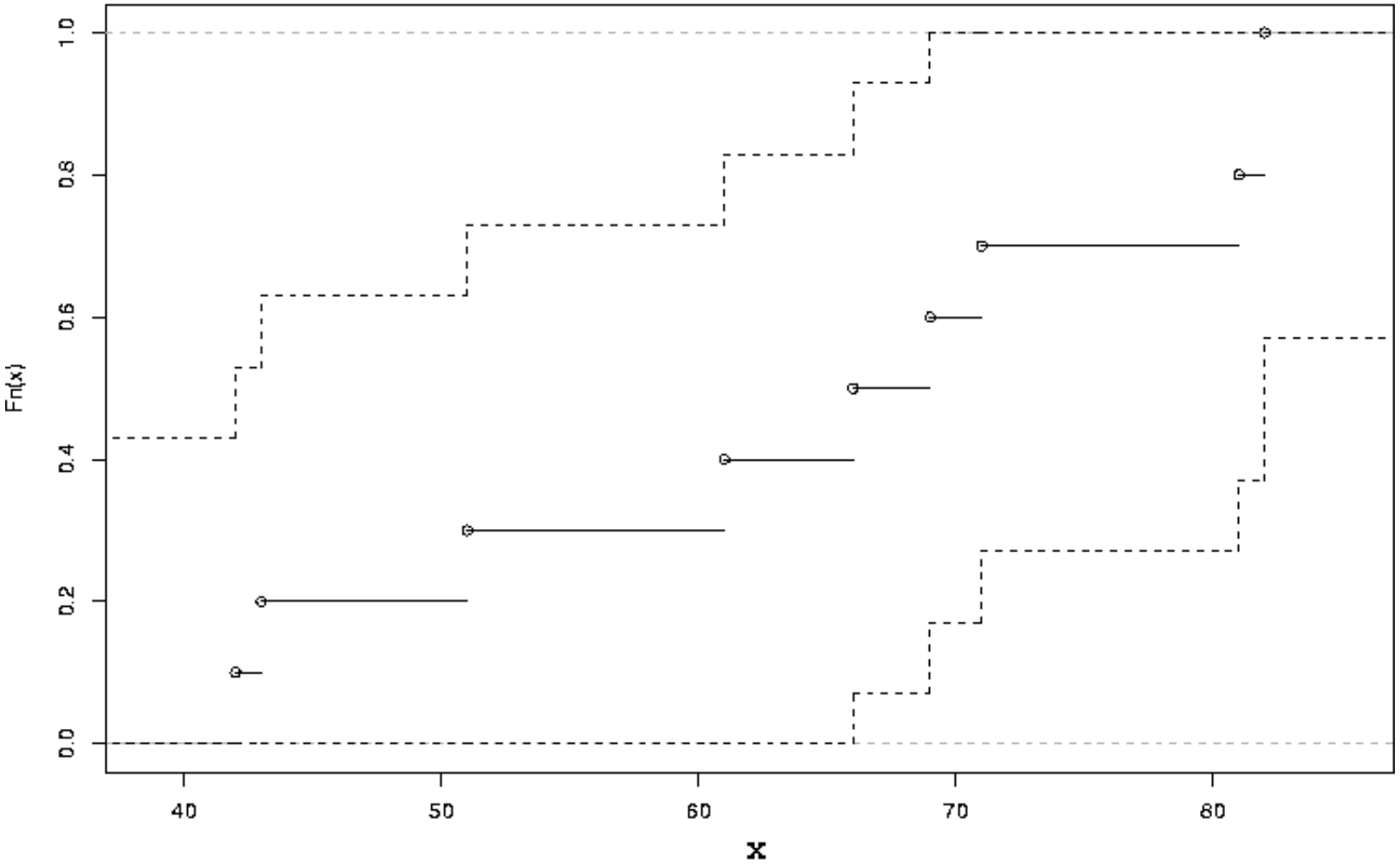
$$H_F(y) = P(D_{n,F} \leq y). \quad 1 - H_F(d_n(\alpha)) = \alpha$$

Cramér-von Mises: $\int (F_n(x) - F(x))^2 \, dF(x)$

Anderson - Darling: $\int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x)$

- These statistics are distribution free if F is continuous & univariate, *i.e.*, $H_F = H$ does not depend on F.
- *No longer distribution free if either F is not univariate or parameters of F are estimated.*

K-S Confidence bands



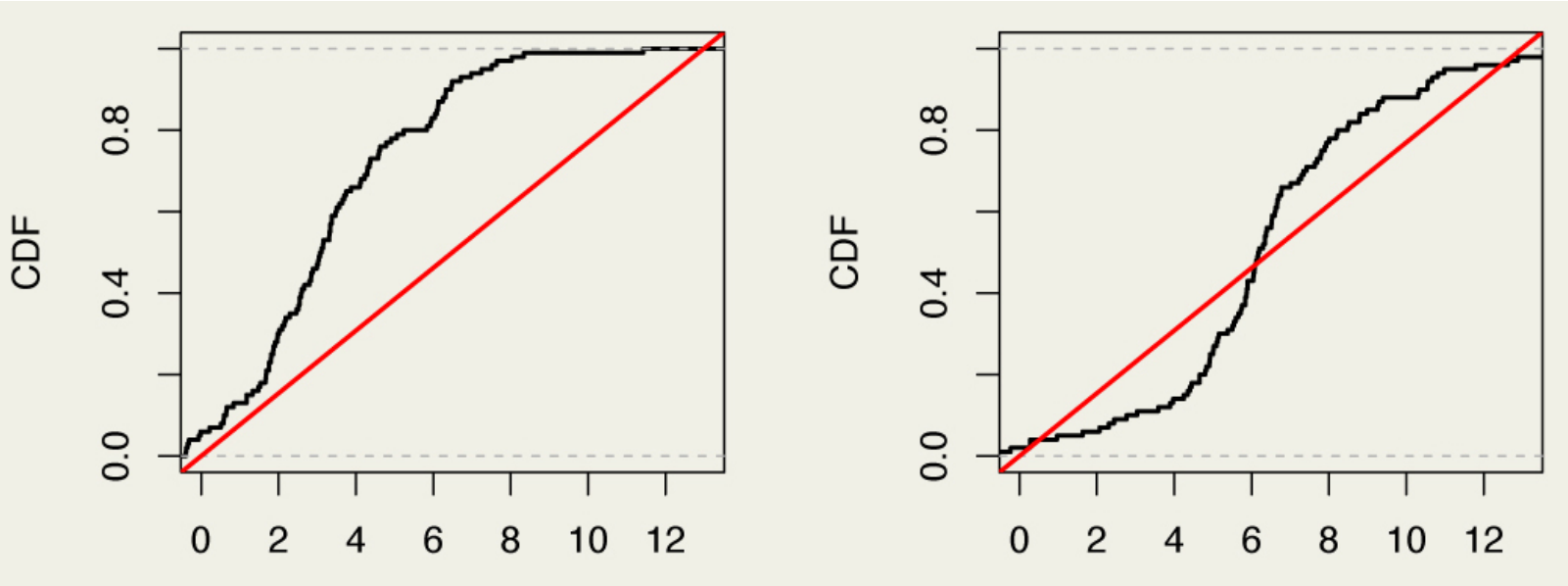
$$F = F_n \pm d_n(\alpha)$$

Kolmogorov-Smirnov and Anderson-Darling Statistics

The K-S statistic efficiently detects differences in global shapes, but not small scale effects or differences near the tails.

The Anderson-Darling (tail-weighted Cramer-von Mises) statistic is more sensitive.

$$KS_n = \sqrt{n} \sup_x |F_n(x) - F(x)| \qquad AD_n = n \int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$



Uses and Misuses of Kolmogorov-Smirnov

- The K-S statistic is used in hundreds of astronomical papers/yr, but often incorrectly or with less efficiency than an alternative statistic.
- *EDF based fitting requires little or no probability distributional assumptions such as Gaussianity or Poisson structure*.
- The 1-sample K-S test (data vs. model comparison) is distribution-free only in 1-dimension and when the model parameters are not derived from the dataset.

  Some astronomers use them incorrectly. – H.W. Lilliefors (1967)

Multidimensional Case

K-S fails in 2-dimensional case. – Paul B. Simpson (1951)

$$F(x, y) = ax^2y + (1 - a)y^2x, \qquad 0 < x, y < 1$$

$(X_1, Y_1) \sim F$. F_1 denotes the EDF of (X_1, Y_1)

$$\begin{aligned} P(|F_1(x, y) - F(x, y)| < .72, \text{ for all } x, y) \\ > .065 \text{ if } a = 0, \quad (F(x, y) = y^2x) \\ < .058 \text{ if } a = .5, \quad (F(x, y) = \frac{1}{2}xy(x + y)) \end{aligned}$$

Numerical Recipe's treatment of a 2-dim K-S test is mathematically invalid.

In the multi-dimension case or when the model parameters are estimated, then the probabilities need to be obtained from **bootstrap** resampling.

Monte Carlo simulation

- Astronomers have often used *Monte Carlo methods* to simulate datasets from power law, uniform, or Gaussian populations. While helpful in some cases, this does not avoid the assumptions of a simple underlying distribution.
- Instead, what if we take the observed data as hypothetical 'population' and use Monte Carlo simulation on it.
- Can simulate many 'datasets' and, each of these can be analyzed in the same way to see how the estimates depend on plausible random variations in the data.

  (No new/additional costly observations). This is exactly the underlying principle behind the Bootstrap procedure.

What is Bootstrap?

- Bootstrap (a resampling procedure) is a Monte Carlo method of simulating 'datasets' from an observed/given data, without any assumption on the underlying population.
- Resampling the original data preserves (adaptively) whatever distributions are truly present, including selection effects such as truncation (flux limits or saturation).
- Bootstrap helps evaluate statistical properties using data rather than an assumed Gaussian or power law or other distributions.
- Bootstrap procedures are supported by solid theoretical foundations.

Bootstrap Procedure

$\boldsymbol{X} = (X_1, \ldots, X_n)$ - a sample from F

$\boldsymbol{X}^* = (X_1^*, \ldots, X_n^*)$ - a simple random sample from the data

$\hat{\theta} = h_n(X_1, \ldots, X_n)$ is an estimator of θ

$\theta^* = h_n(X_1^*, \ldots, X_n^*)$ is based on X_i^*

Examples:

$$\hat{\theta} = \bar{X}_n, \qquad \theta^* = \bar{X}_n^*$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \qquad \theta^* = \frac{1}{n}\sum_{i=1}^{n}(X_i^* - \bar{X}_n^*)^2$$

$\theta^* - \hat{\theta}$ behaves like $\hat{\theta} - \theta$

Sampling & Bootstrap Distributions

Units free/standardized statistic: $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

Its sampling distribution G_n is given by

$$G_n(x) = \mathrm{P}(\sqrt{n}(\bar{X} - \mu)/\sigma \leq x).$$

For a given data $\mathbf{X}$, the bootstrap version of the standardized statistic is

$$\frac{\bar{X}^* - \bar{X}}{s_n/\sqrt{n}}, \quad \text{where} \quad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The bootstrap distribution

$$G_B(x) = \mathrm{P}^*(\sqrt{n}(\bar{X}^* - \bar{X})/s_n \leq x|\mathbf{X})$$

is completely known and $G_n \approx G_B$.

Construction of Bootstrap Distribution

$M =$ total number of bootstrap samples possible

$$
\begin{array}{cc}
X_1^{*(1)}, \ldots, X_n^{*(1)} & r_1 = \sqrt{n}(\bar{X}^{*(1)} - \bar{X})/s_n \\
X_1^{*(2)}, \ldots, X_n^{*(2)} & r_2 = \sqrt{n}(\bar{X}^{*(2)} - \bar{X})/s_n \\
\ddots \qquad \ddots & \ddots \qquad \ddots \\
X_1^{*(M)}, \ldots, X_n^{*(M)} & r_M = \sqrt{n}(\bar{X}^{*(M)} - \bar{X})/s_n
\end{array}
$$

Empirical distribution based on $r_1, \ldots, r_M$ gives G_B:

$$G_B(x) = \frac{1}{M}\#(1 \leq i \leq M\colon\ r_i \leq x).$$

Confidence Interval for the mean

For $n = 10$ data points, $M = 10^{10}$ ten billion

$N \sim n(\log n)^2$ bootstrap replications suffice
N is much smaller than n^n.

– Babu and Singh (1983) Ann. Stat.

Compute $r_{i_j} = \sqrt{n}(\bar{X}^{*(i_j)} - \bar{X})/s_n$ for N bootstrap samples

Arrange them in increasing order

$v_1 < v_2 < \cdots < v_N \qquad k = [0.05N],\ m = [0.95N]$

90% Confidence Interval for μ is

$$\bar{X} - v_m \frac{s_n}{\sqrt{n}} \leq \mu < \bar{X} - v_k \frac{s_n}{\sqrt{n}}$$

Bootstrap at its best

Pearson's correlation coefficient and its bootstrap version

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i Y_i - \bar{X}\bar{Y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)}}$$

$$\rho^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i^* Y_i^* - \bar{X}_n^* \bar{Y}_n^*)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^* - \bar{X}_n^*)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \bar{Y}_n^*)^2\right)}}$$

Smooth Functional Model

$$\hat{\rho} = H(\bar{\mathbf{Z}}), \quad \text{where} \quad \mathbf{Z}_i = (X_i Y_i, X_i^2, Y_i^2, X_i, Y_i)$$

$$H(a_1, a_2, a_3, a_4, a_5) = \frac{(a_1 - a_4 a_5)}{\sqrt{((a_2 - a_4^2)(a_3 - a_5^2))}}$$

$$\rho^* = H(\bar{\mathbf{Z}}^*), \quad \text{where} \quad \mathbf{Z}_i^* = (X_i^* Y_i^*, X_i^{*2}, Y_i^{*2}, X_i^*, Y_i^*)$$

Smooth Functional Model: General case

H is a smooth function and $\mathbf{Z}_1$ is a random vector.

$\hat{\theta} = H(\bar{\mathbf{Z}})$ is an estimator of the parameter $\theta = H(\mathrm{E}(\mathbf{Z}_1))$

Division (normalization) of $\sqrt{n}(H(\bar{\mathbf{Z}}) - H(\mathrm{E}(\mathbf{Z}_1))$ by its standard deviation makes them units free.

Studentization, if estimates of standard deviations are used.

Under some regularity conditions Bootstrap distribution gives a very good approximation to the sampling distribution of such normalized/Studentized statistics.

– Babu and Singh (1983) Ann. Stat.
– Babu and Singh (1984) Sankhyā
– Singh and Babu (1990) Scand J. Stat.

When does bootstrap work well

- Sample Means
- Sample Variances
- Central and Non-central t-statistics
  (with possibly non-normal populations)
- Sample Coefficient of Variation
- Maximum Likelihood Estimators
- Least Squares Estimators
- Correlation Coefficients
- Regression Coefficients
- Smooth transforms of these statistics

When does Bootstrap fail

- $\hat{\theta} = \max_{1 \le i \le n} X_i$ Non-smooth estimator

  – Bickel and Freedman (1981) Ann. Stat.

- $\hat{\theta} = \bar{X}$ and $\mathrm{E}X_1^2 = \infty$ Heavy tails

  – Babu (1984) Sankhyā
  – Athreya (1987) Ann. Stat.

Non-independent case

$X_1, \ldots X_n$ are identically distributed but not independent

- Straight forward bootstrap does not work in the dependent case. Variances of sums of random variables do not match.
- A clear knowledge of the dependent structure is needed to replicate resampling procedure.
- Classical bootstrap fails in the case of Time Series data.
- If the process is auto-regressive or moving-average one can replicate resampling procedure.
- In the general time-series case the *moving block bootstrap* is suggested.

Moving Block Bootstrap

$X_1, \cdots, X_n$ is a stationary sequence.

1. The sequence is split into overlapping blocks $B_1, \cdots, B_{n-b+1}$, of length b, where B_j consists of b consecutive observations starting from X_j, i.e., $B_j = \{X_j, X_{j+1}, \cdots, X_{j+b-1}\}$.
   Observation 1 to b will be block 1, observation 2 to b+1 will be block 2 etc.
2. From these n-b+1 blocks, n/b blocks will be drawn at random with replacement.
3. Align these n/b blocks in the order they were picked.

This bootstrap procedure works with dependent data.
By construction, the resampled data will not be stationary.

Varying randomly the block length can avoid this problem.
However, the moving block bootstrap is still to be preferred.

– Lahiri (1999) Annals of Statistics

Nonparametric and Parametric Bootstrap

Simple random sampling from data is equivalent to drawing a set of i.i.d. random variables from the empirical distribution.
This is Nonparametric Bootstrap.

Parametric Bootstrap if $X_1^*, \ldots, X_n^*$ are i.i.d. r.v. from $\hat{H}_n$, an estimator of F based on data $(X_1, \ldots, X_n)$.

Example of Parametric Bootstrap:

$X_1, \ldots, X_n \ i.i.d. \sim N(\mu, \sigma^2)$

$X_1^*, \ldots, X_n^* \ i.i.d. \sim N(\bar{X}_n, s_n^2); \quad s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

$N(\bar{X}_n, s_n^2)$ is a good estimator of the distribution $N(\mu, \sigma^2)$

Goodness of Fit when parameters are estimated

$X_1, \ldots, X_n$ sample from $F \in \{F(.;\theta) : \theta \in \Theta\}$ – a family of continuous distributions. Θ is p-dimensional.

$X_1^*, \ldots, X_n^*$ sample generated from $F(.;\hat{\theta}_n)$

The bootstrap version of K-S statistic

$$\sup_x |F_n(x) - F(x;\hat{\theta}_n)| \quad \text{is} \quad \sup_x |F_n^*(x) - F(x;\hat{\theta}_n^*)|.$$

In XSPEC package, the parametric bootstrap is command FAKEIT, which makes Monte Carlo simulation of specified spectral model.

In Gaussian case $\hat{\theta}_n^* = (\bar{X}_n^*, s_n^{*2})$.

Numerical Recipes describes a parametric bootstrap (random sampling of a specified pdf) as the 'transformation method' of generating random deviates.

Summary

- EDF based fitting requires little or no probability distributional assumptions such as Gaussianity or Poisson structure.
- K-S goodness of fit is often better than Chi-square test.
- K-S cannot handle heteroscadastic errors
- Anderson-Darling is better in handling the tail part of the distributions.
- K-S probabilities are incorrect if the model parameters are estimated from the same data.
- K-S does not work in more than one dimension.
- Bootstrap helps in the last two cases.

Bootstrap References

- Babu, G. J., and Rao, C. R. (1993). Bootstrap methodology. In *Computational statistics*, Handbook of Statistics **9**, C. R. Rao (Ed.), North-Holland, Amsterdam, 627-659.
- Babu, G. J., and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā*, **66**, no. 1, 63-74.
- Michael R. Chernick (2007). *Bootstrap Methods - A guide for Practitioners and Researchers*, (2nd Ed.) Wiley Inter-Science.
- Michael R. Chernick and Robert A. LaBudde (2011) *An Introduction to Bootstrap Methods with Applications to R*, Wiley.
- Abdelhak M. Zoubir and D. Robert Iskander (2004) *Bootstrap Techniques for Signal Processing*, Cambridge Univ Press.

  A handbook on 'bootstrap' for engineers to analyze complicated data with little or no model assumptions. Includes applications to radar and sonar signal processing.