

Statistics: Basic Concepts

Aneta Siemiginowska
Harvard-Smithsonian Center for Astrophysics

OUTLINE

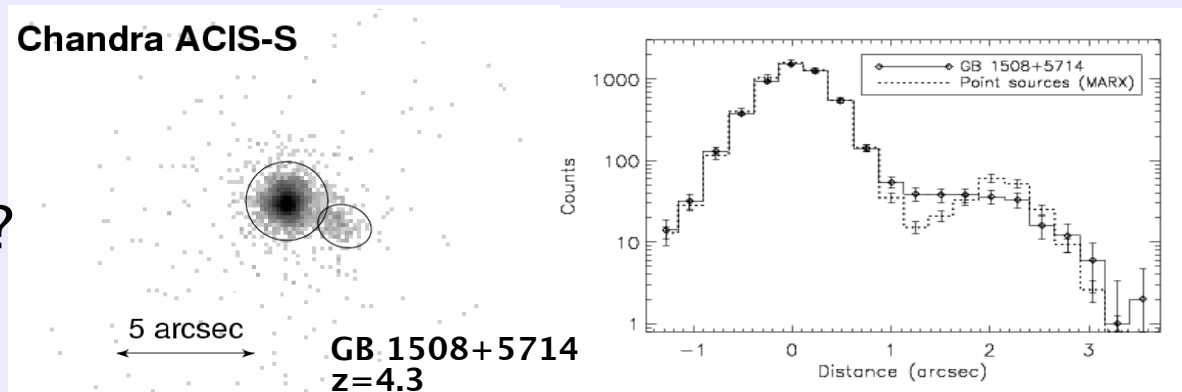
- Motivation: why do we need statistics?
- Probabilities/Distributions
- Poisson Likelihood
- Parameter Estimation
- Statistical Issues

Why do we need Statistics?

- How do we take decisions in Science?
Tools: instruments, data collections, reduction, classifications – tools and techniques
Decisions: is this hypothesis correct? Why not? Are these data consistent with other data? Do we get an answer to our question? Do we need more data?
- Comparison to decide :
 - Describe properties of an object or sample:

Example:

Is a faint extension a jet or a point source?



Siemiginowska et al (2003)

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)

Statistic is a quantity that summarizes data

Statistics are combinations of data that do not depend on unknown parameters:
Mean, averages from multiple experiments
etc.

=> Astronomers cannot avoid Statistics



Probability

Numerical formalization of our degree of belief.



$$\frac{\text{Number of favorable events}}{\text{Total number of events}}$$

Laplace principle of indifference:

All events have equal probability

Example 1:

1/6 is the probability of throwing a 6 with 1 roll of the dice **BUT** the dice can be biased! => need to calculate the probability of each face

Example 2:

Use data to calculate probability, thus the probability of a cloudy observing run:

$$\frac{\text{number of cloudy nights last year}}{365 \text{ days}}$$

Issues:

- limited data
- not all nights are *equally likely* to be cloudy

Properties of Probability

Formalize the “measure of belief”:

A,B,C – three events and we need to measure how strongly we think each is likely to happen and apply the rule:

If A is more likely than B, and B is more likely than C, then A is more likely than C.

Kolmogorov axioms - Foundation of the Theory of Probability

- Any random event A has a probability **prob(A) between 0 and 1**
- The sure event **prob(A) = 1**
- If A and B are exclusive ($A \cap B = \emptyset$), disjoint events then **prob(A \cup B) = prob(A) + prob(B)**

Conditionality and Independence

A and B events are **independent** if the probability of one is unaffected by what we know about the other:

$$\text{prob}(A \text{ and } B) = \text{prob}(A)\text{prob}(B)$$

If the probability of A depends on what we know about B
A given B \Rightarrow **conditional probability**

$$\text{prob}(A|B) = \frac{\text{prob}(A \text{ and } B)}{\text{prob}(B)}$$

If A and B are independent \Rightarrow $\text{prob}(A|B) = \text{prob}(A)$

If there are several possibilities for event B (B_1, B_2, \dots)

$$\text{prob}(A) = \sum \text{prob}(A|B_i) \text{prob}(B_i)$$

A – parameter of interest

B_i – not of interest, instrumental parameters, background

$\text{prob}(B_i)$ – if known we can sum (or integrate) – **Marginalize**

Bayes' Theorem

Bayes' Theorem is derived by equating:
 $\text{prob}(A \text{ and } B) = \text{prob}(B \text{ and } A)$

$$\text{prob}(B|A) = \frac{\text{prob}(A|B) \text{prob}(B)}{\text{prob}(A)}$$

Gives the Rule for induction:

the data, the event A, are succeeding B, the state of belief preceding the experiment.

$\text{prob}(B)$ – **prior probability** which will be modified by experience

$\text{prob}(A|B)$ – **likelihood**

$\text{prob}(B|A)$ – **posterior probability** – the state of belief after the data have been analyzed

$\text{prob}(A)$ – normalization

Example

A box with colored balls:
what is the content of the box?

$$\text{prob}(\text{content of the box} \mid \text{data}) \propto \text{prob}(\text{data} \mid \text{content of the box})$$

Experiment:

N red balls

M white balls

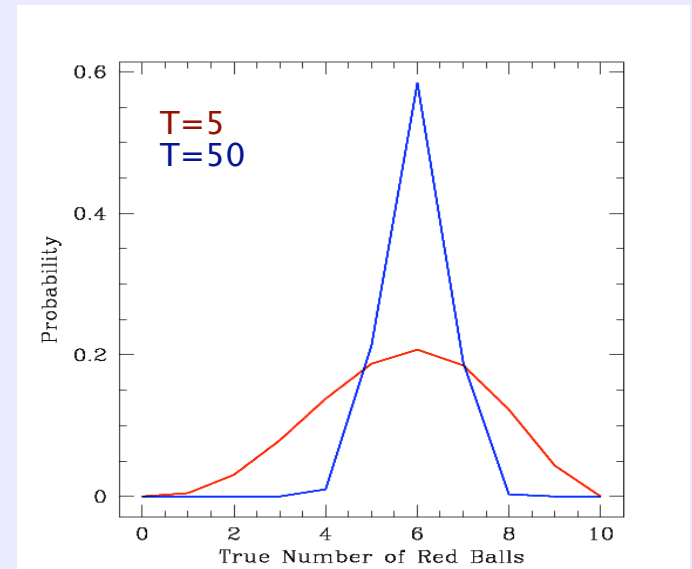
N+M = 10 total, known

Draw 5 times (putting back) (T) and
get 3 red balls (R)

How many red balls are in the box?

$$\text{Model (our hypothesis)} \Rightarrow \text{prob}(R) = \frac{N}{N+M}$$

$$\text{Likelihood} = \binom{T}{R} \text{prob}(R)^R \text{prob}(M)^{T-R}$$



Example

From a paper by Martinez-Sansigre et al published in Aug 4, 2005 issue of *Nature*

What is the fraction of the unobscured quasars?

Use new Spitzer observations

q - quasar fraction

$$q = \frac{\text{Type-1 quasars}}{\text{Type-1} + \text{Type-2}} = \frac{N1}{N1+N2}$$

$\langle N1 \rangle$ - number of Type-1 qso

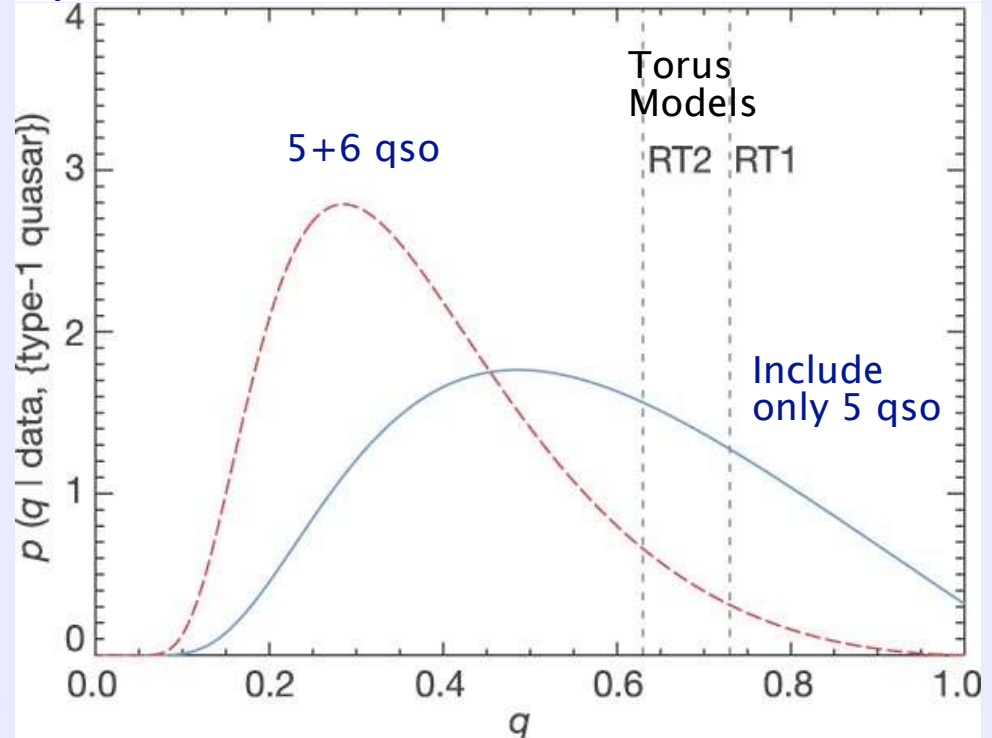
$\langle N2 \rangle$ - number of Type-2 qso

1/ take Poisson likelihood with the mean $\langle N2 \rangle = (1-q)\langle N1 \rangle / q$

2/ evaluate likelihood at each q and $N1$

3/ integrate $P(N1|q)P(N1)$ over $N1$

Posterior Probability distribution for the quasar fraction

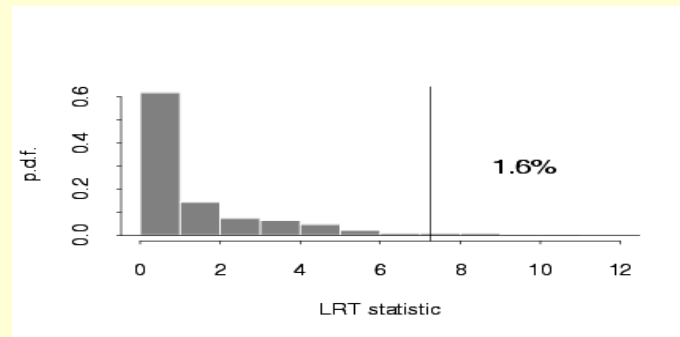


$$p(q|\text{data},\{\text{type-1 qso}\}) = p(\text{data}|q,\{\text{type-1 qso}\})$$

Probability Distributions

Probability is crucial in decision process:

Example:



Limited data yields only partial idea about the line width in the spectrum. We can only assign the probability to the range of the line width roughly matching this parameter. We decide on the presence of the line by calculating the probability.

Definitions

- **Random variable:** a variable which can take on different numerical values, corresponding to different experimental outcomes.
 - Example: a binned datum D_i , which can have different values even when an experiment is repeated exactly.
- **Statistic:** a function of random variables.
 - Example: a datum D_i , or a population mean

$$(\mu = [\sum_{i=1}^N D_i] / N)$$

- **Probability sampling distribution:** the normalized distribution from which a statistic is sampled. Such a distribution is commonly denoted $p(X|Y)$, “the probability of outcome X given condition(s) Y ,” or sometimes just $\mathbf{p}(X)$. Note that in the special case of the Gaussian (or normal) distribution, $\mathbf{p}(X)$ may be written as $N(\mu, \sigma^2)$, where μ is the Gaussian mean, and σ^2 is its variance.

The Poisson Distribution

**Collecting X-ray data => Counting individual photons
=> Sampling from Poisson distribution**

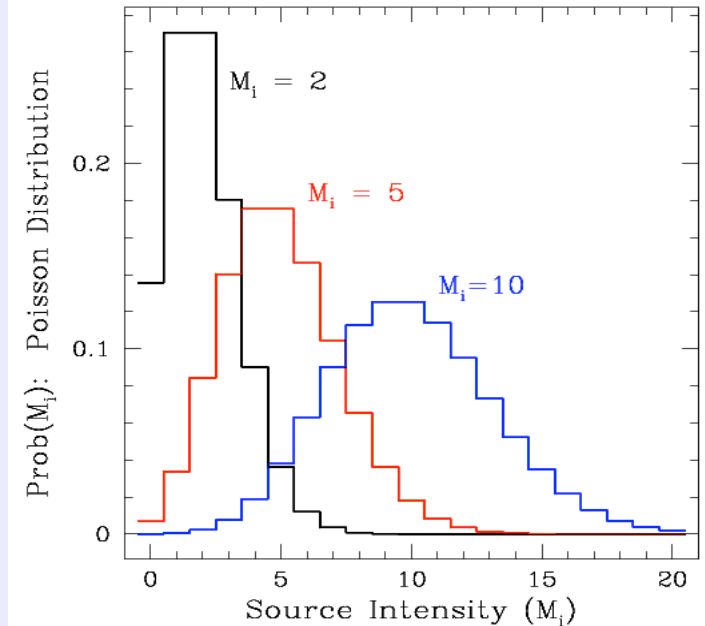
The discrete Poisson distribution:

$$\text{prob}(D_i) = p(D_i | M_i) = \frac{M_i^{D_i}}{D_i!} e^{-M_i}$$

probability of finding D_i events (**counts**) in bin i (**energy rage**) of dataset D (**spectrum**) in a given length of time (exposure time), if the events occur independently at a constant rate M_i (**source intensity**).

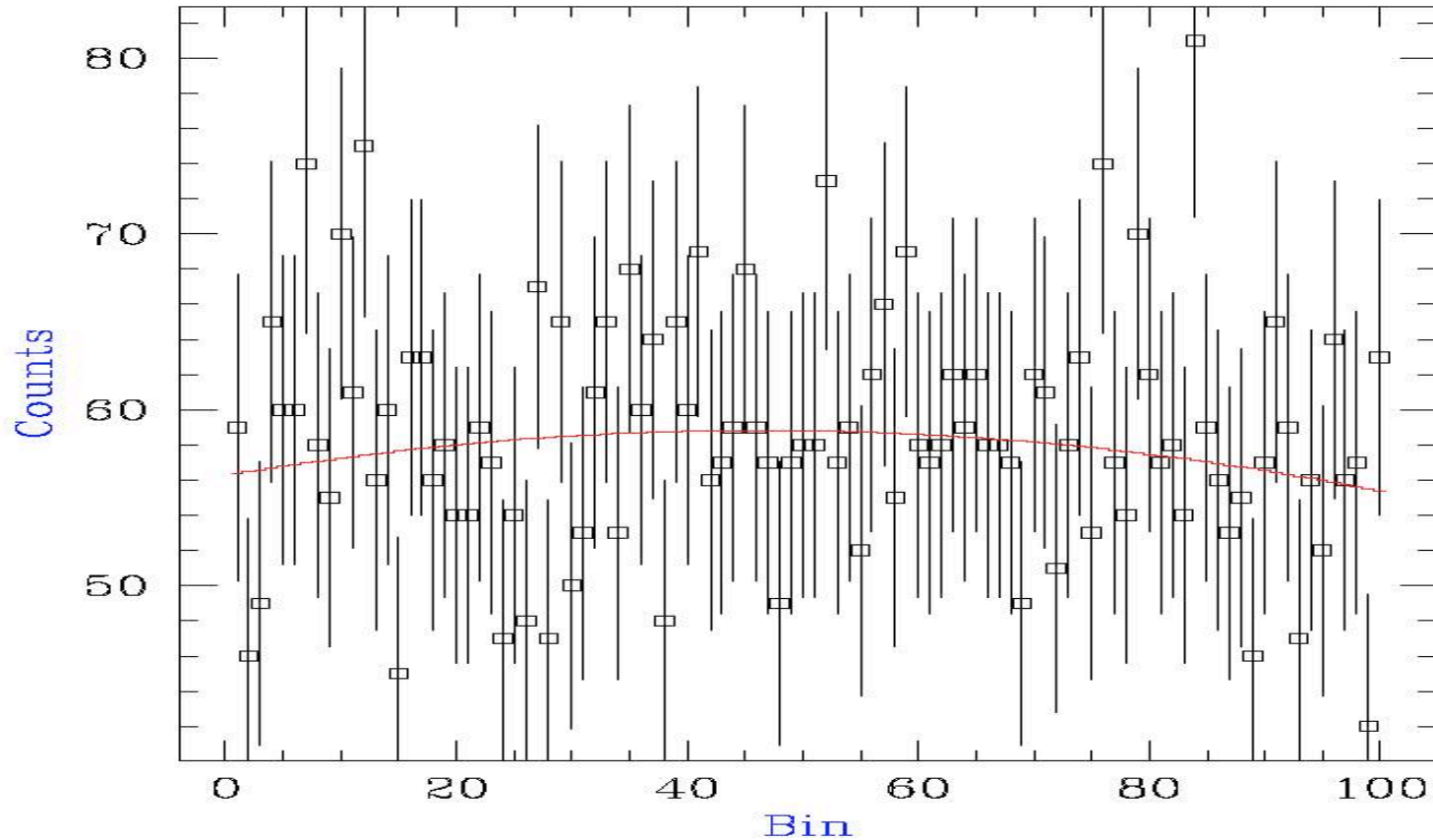
Things to remember:

- Mean $\mu = E[D_i] = M_i$
- Variance: $V[D_i] = M_i$
- $\text{cov}[D_{i_1}, D_{i_2}] = 0 \Rightarrow$ independent
- the sum of n Poisson-distributed variables is itself Poisson-distributed with variance: $\sum_{i=1}^n M_i$



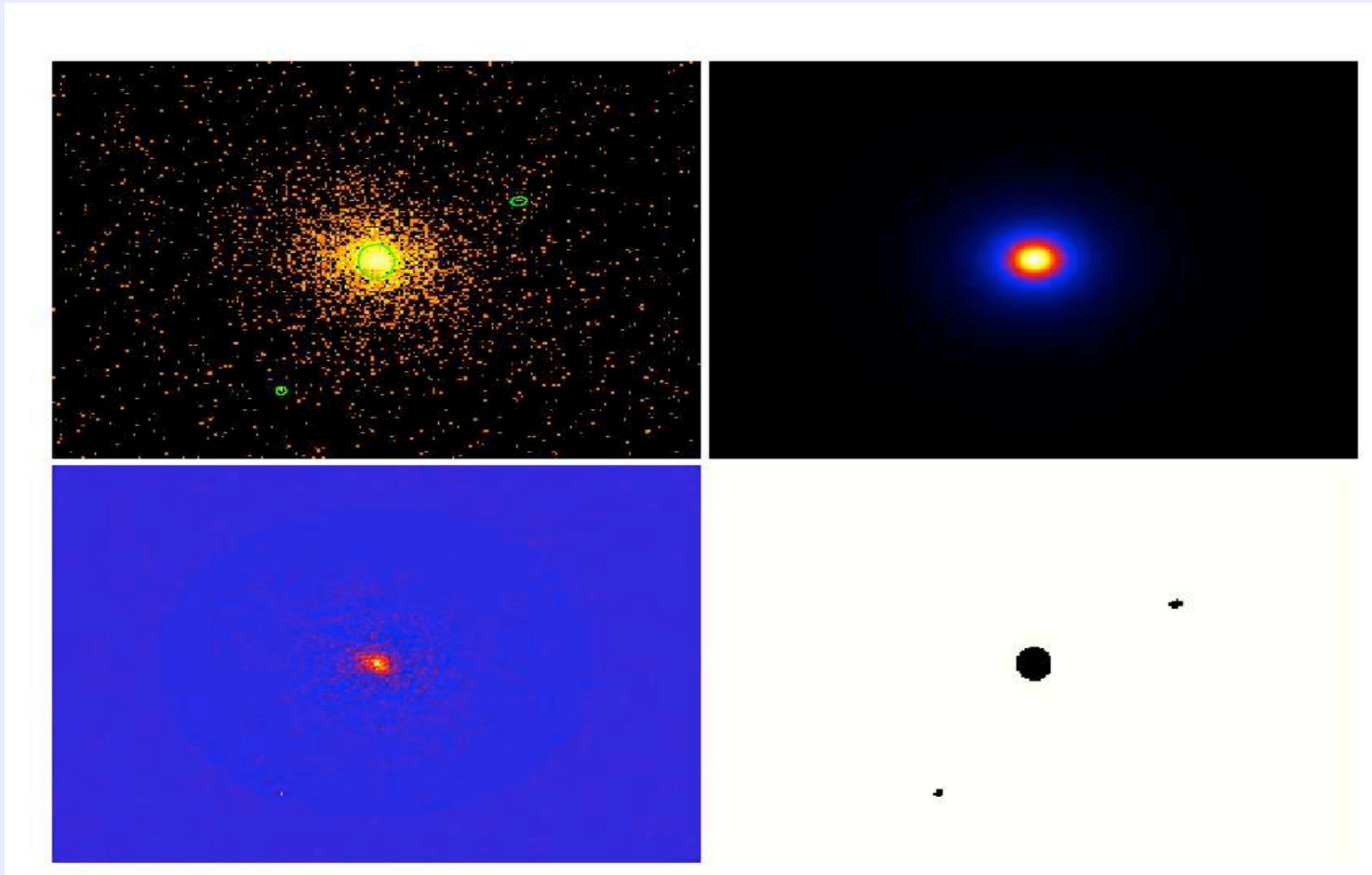
As $M_i \rightarrow \infty$ Poisson distribution converges to Gaussian distribution $N(\mu = M_i; \sigma^2 = M_i)$

Example:



Integer counts spectrum sampled from a constant amplitude model with mean $\mu = 60$ counts, and fit with a parabolic model.

Example2



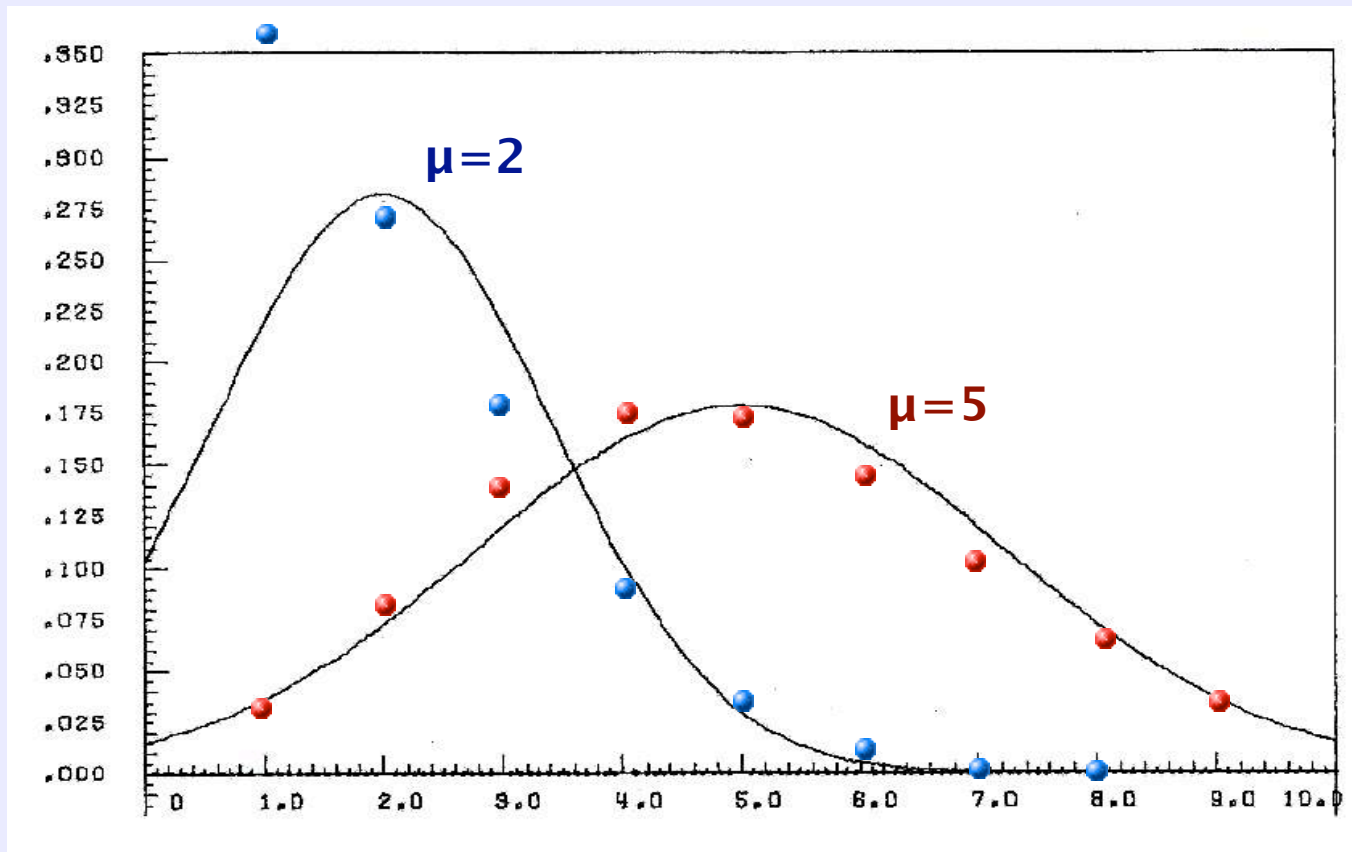
Example of a two-dimensional integer counts spectrum. **Top Left:** *Chandra* ACIS-S data of X-ray cluster MS 2137.3–2353, with *ds9* source regions superimposed.

Top Right: Best-fit of a two-dimensional beta model to the filtered data.

Bottom Left: Residuals (in units of σ) of the best fit.

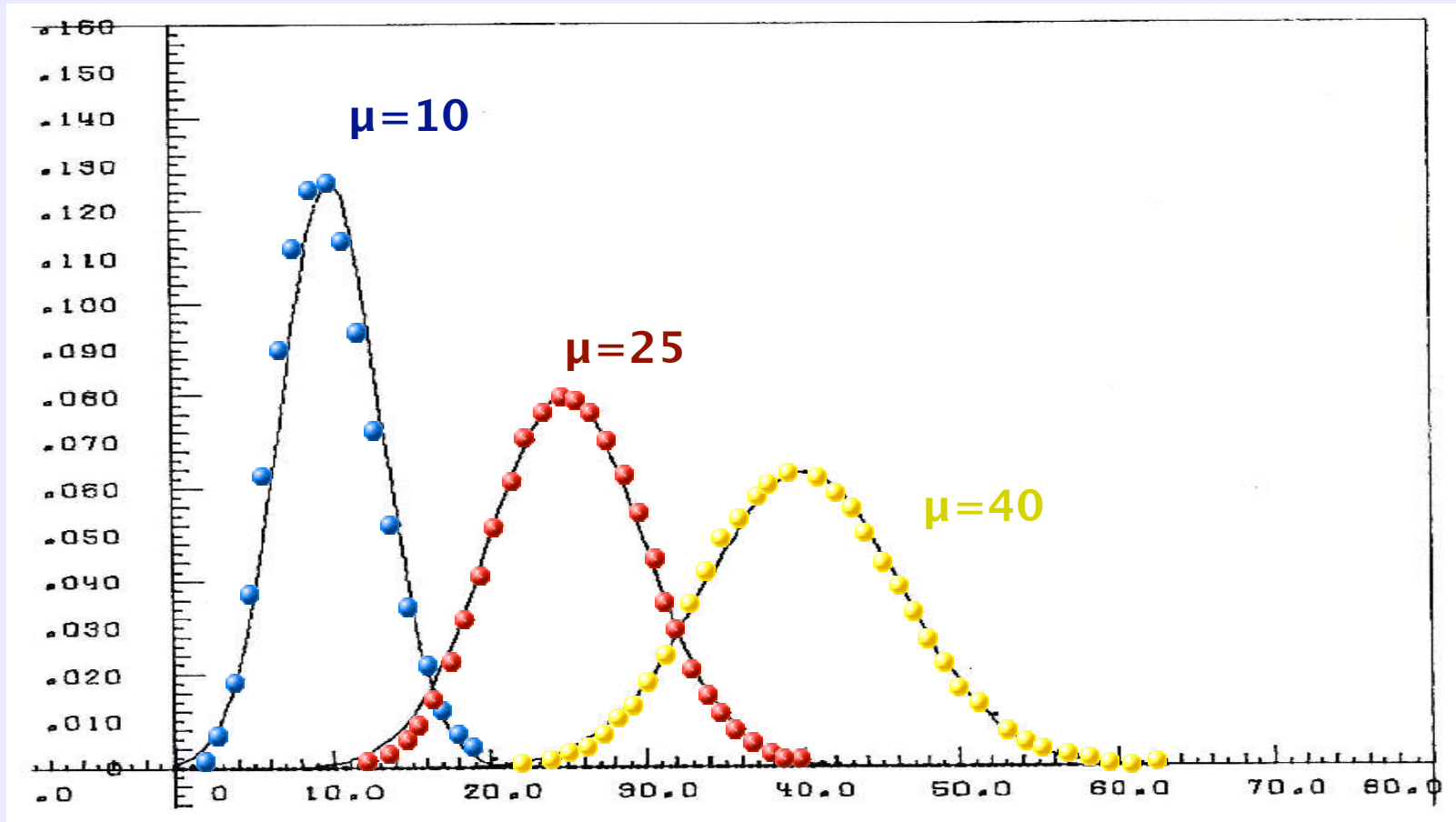
Bottom Right: The applied filter; the data within the ovals were excluded from the fit.

Poisson vs. Gaussian Distributions - Low Number of Counts



Comparison of Poisson distributions (dotted) of mean $\mu = 2$ and 5 with normal distributions of the same mean and variance (Eadie *et al.* 1971, p. 50).

Poisson vs. Gaussian Distributions



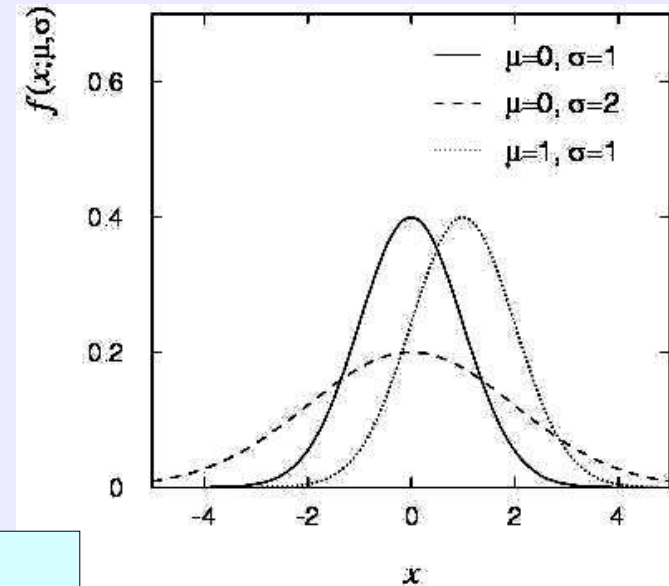
Comparison of Poisson distributions (dotted) of mean $\mu = 10, 25$ and 40 with normal distributions of the same mean and variance (Eadie *et al.* 1971, p. 50).

Gaussian Distribution

For large $\mu \rightarrow \infty$ Poisson (and the Binomial, large T) distributions converge to Gaussian (normal) distributions.

$$\text{prob}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2]$$

Mean - μ
Variance - σ^2



Note: Importance of the Tails!

$\pm 2\sigma$ range covers **95.45%** of the area, so 2σ result has less than **5%** chance of occurring by chance, but because of the error estimates are difficult this is not the acceptable result. Usually **3σ** or **10σ** have to be quoted and the convergence to Gaussian fastest in the center than in the tails!

Central Limit Theorem

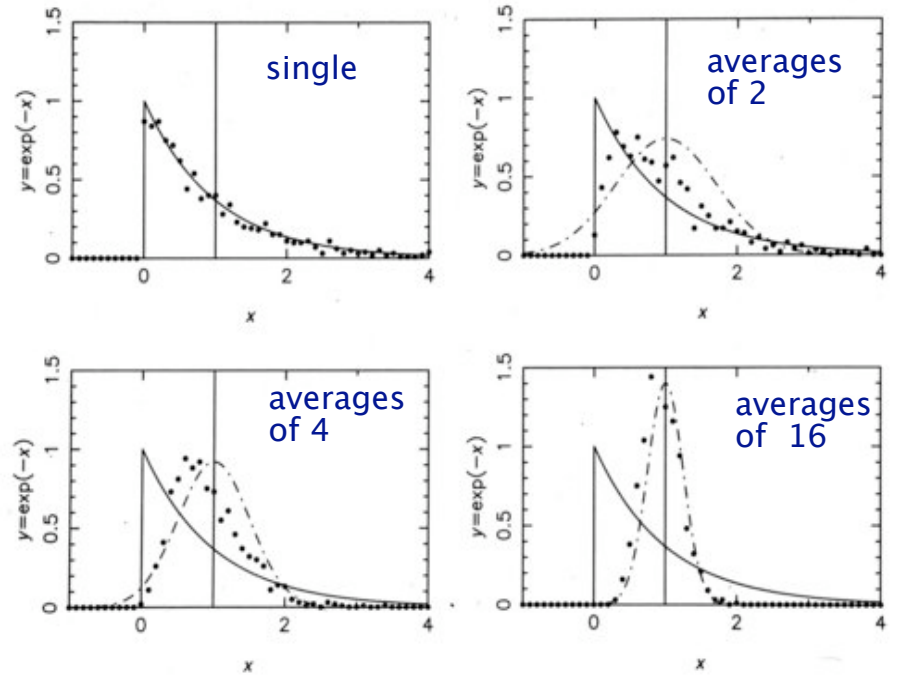
The true importance of the Gaussian distribution

Form averages M_n from repeated drawing of n samples from a population with finite mean μ and variance σ^2

$$\frac{(M_n - \mu)}{\sigma/\sqrt{n}}$$

=> Gaussian Distribution
as $n \rightarrow \infty$

$$\mu = 0, \sigma^2 = 1$$



200 y values drawn from exp(-x) function

Bayesian vs. Classical

Example:

$$D = 8.5 \mp 0.1 \text{ Mpc}$$

Does not describe probability that a true value is between 8.4 and 8.6.

We **assume** that a Gaussian distribution applies and knowing the distribution of errors we can make probabilistic statements.

Classical Approach:

Assuming the true distance D_0 then D is normally distributed around D_0 with a standard deviation of 0.1. Repeating measurement will yield many estimates of distance D which all scatter around true D_0 .

Assume the thing (distance) we want to know and tell us how the data will behave.

Bayesian Approach:

Deduce directly the probability distribution of D_0 from the data.
Assumes the data and tell us the thing we want to know. No repetition of experiment.

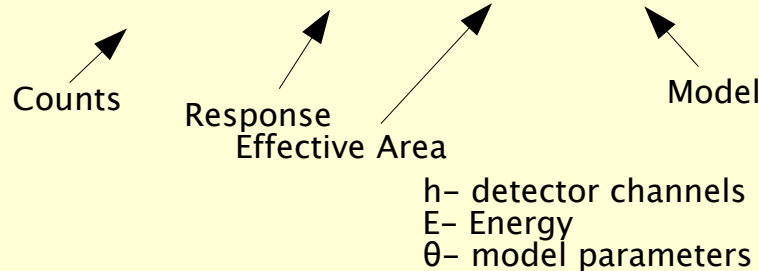
What do we really do?

Example:

I've observed my source, reduce the data and finally got my X-ray spectrum – what do I do now? How can I find out what does the spectrum tell me about the physics of my source?

Run **XSPEC** or **Sherpa**! But what do those programs really do?

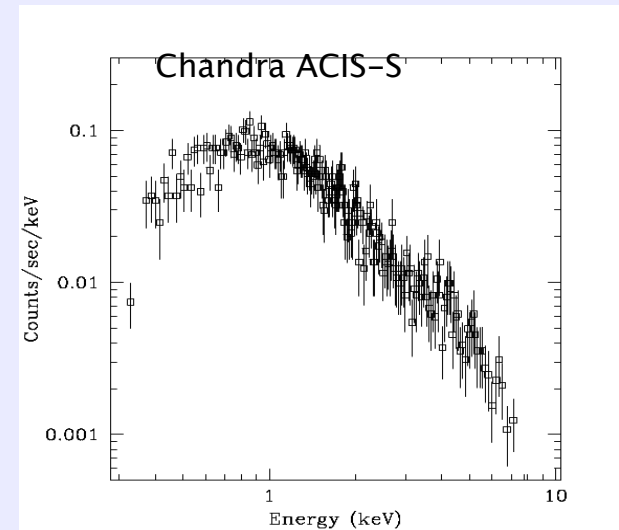
Fit the data => $C(h) = \int R(E, h) A(E) M(E, \theta) dE$



Assume a model and look for the best model parameters which describes the observed spectrum.



Need a Parameter Estimator – Statistics



Statistics

(1) Statistics indicating the location of the data:

Average: $\langle X \rangle = (1/N) \sum_i X_i$

Mode: location of the peak in the histogram; the value occurring most frequently

(2) Statistics indicating the scale or amount of scatter:

Mean deviation: $\langle \Delta X \rangle = (1/N) \sum_i |X_i - \langle X \rangle|$

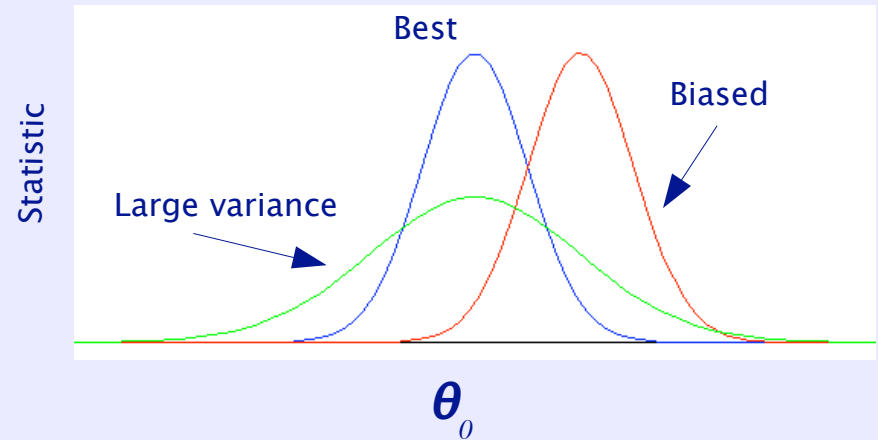
Mean square deviation: $S^2 = (1/N) \sum_i (X_i - \langle X \rangle)^2$

Root Mean Square deviation: $\text{rms} = S$

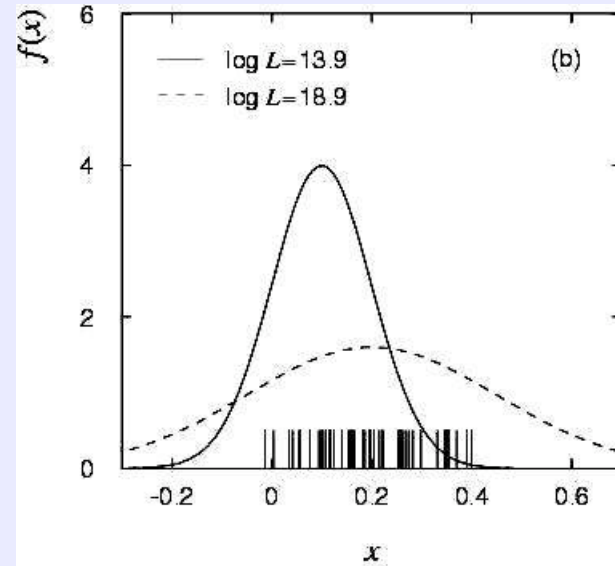
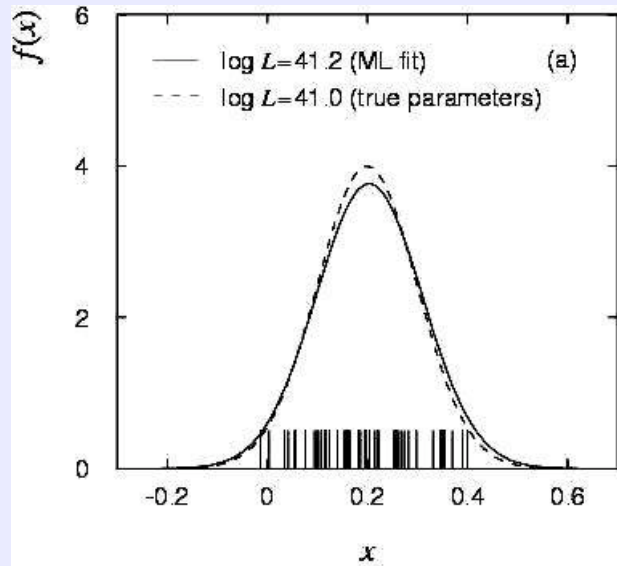
Parameter Estimators

Requirements on Statistics:

- **Unbiased**
 - converge to true value with repeated measurements
- **Robust**
 - less affected by outliers
- **Consistent**
 - true value for a large sample size (Example: rms and Gaussian distribution)
- **Closeness**
 - smallest variations from the truth



If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

Maximum Likelihood: Assessing the Quality of Fit

One can use the Poisson distribution to assess the probability of sampling a datum D_i given a predicted (convolved) model amplitude M_i . Thus to assess the quality of a fit, it is natural to maximize the product of Poisson probabilities in each data bin, *i.e.*, to maximize the Poisson likelihood:

$$L = \prod_i^N L_i = \prod_i^N \frac{M_i^{D_i}}{D_i!} \exp(-M_i) = \prod_i^N p(D_i | M_i)$$

In practice, what is often maximized is the log-likelihood,

$L = \log \mathcal{L}$. A well-known statistic in X-ray astronomy which is related to L is the so-called “Cash statistic”:

$$C \equiv 2 \sum_i^N [M_i - D_i \log M_i] \propto -2L,$$

(Non-) Use of the Poisson Likelihood

In model fits, the Poisson likelihood is not as commonly used as it should be. Some reasons why include:

- a historical aversion to computing factorials;
- the fact the likelihood cannot be used to fit “background subtracted” spectra;
- the fact that negative amplitudes are not allowed (not a bad thing physics abhors negative fluxes!);
- the fact that there is no “goodness of fit” criterion, i.e. there is no easy way to interpret \mathcal{L}_{\max} (however, *cf.* the **CSTAT** statistic); and
- the fact that there is an alternative in the Gaussian limit: the χ^2 statistic.

χ^2 Statistic

Definition: $\chi^2 = \sum_i (D_i - M_i)^2 / M_i$

The χ^2 statistics is **minimized** in the fitting the data, varying the model parameters until the best-fit model parameters are found for the minimum value of the χ^2 statistic

Degrees-of-freedom = $k-1 - N$

N - number of parameters

K - number of spectral bins

Confidence Limits

Essential issue = after the best-fit parameters are found estimate the confidence limits for them. The region of confidence is given by (Avni 1976):

$$\chi^2_{\alpha} = \chi^2_{\min} + \Delta(\nu, \alpha)$$

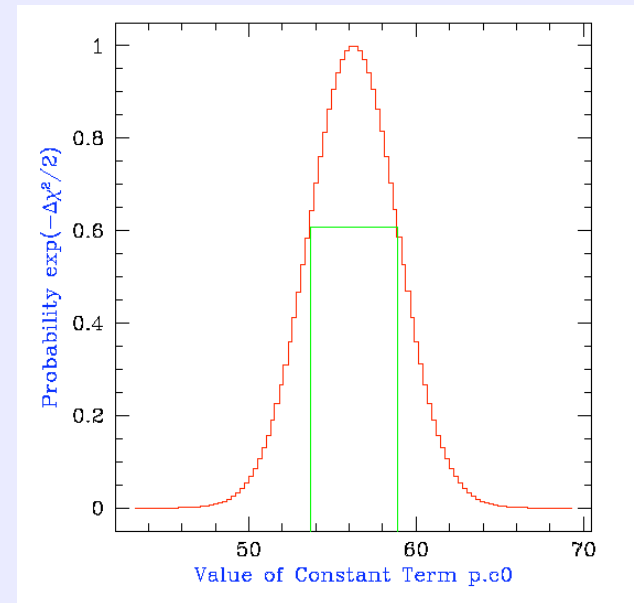
ν – degrees of freedom

α – significance

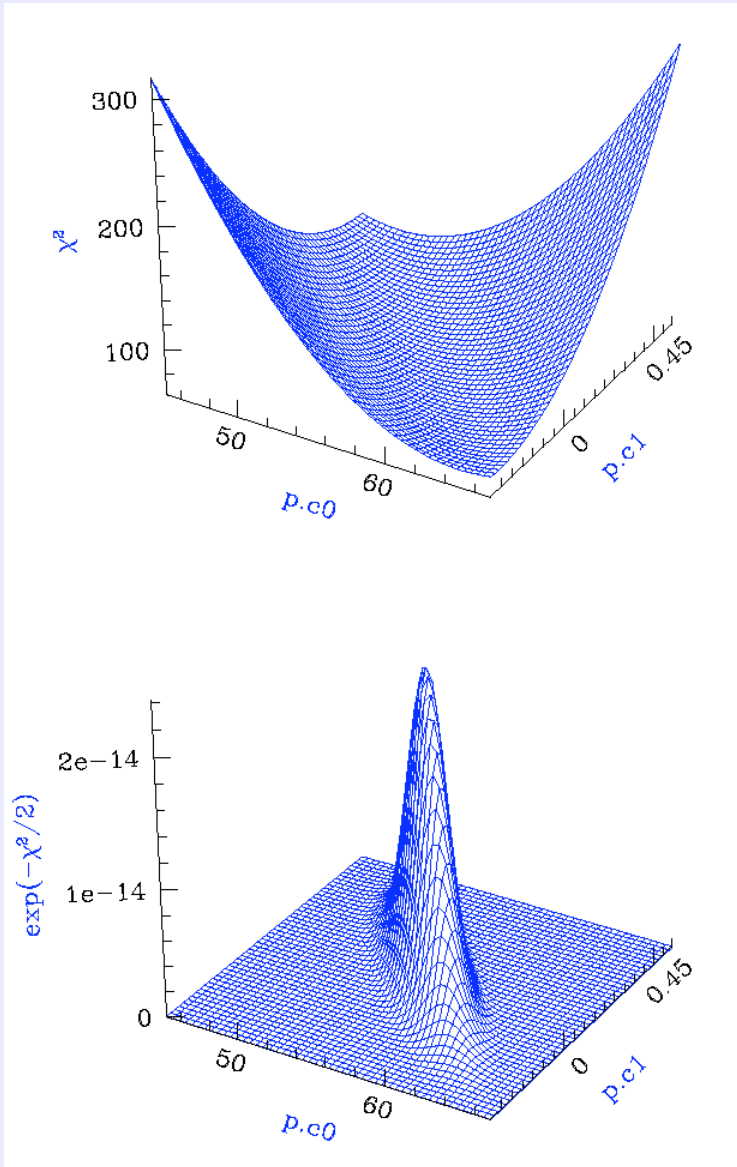
χ^2_{\min} – minimum

Δ - *depends only on the number of parameters involved nor on goodness of fit*

Significance α	Number of parameters		
	1	2	3
0.68	1.00	2.30	3.50
0.90	2.71	4.61	6.25
0.99	6.63	9.21	11.30



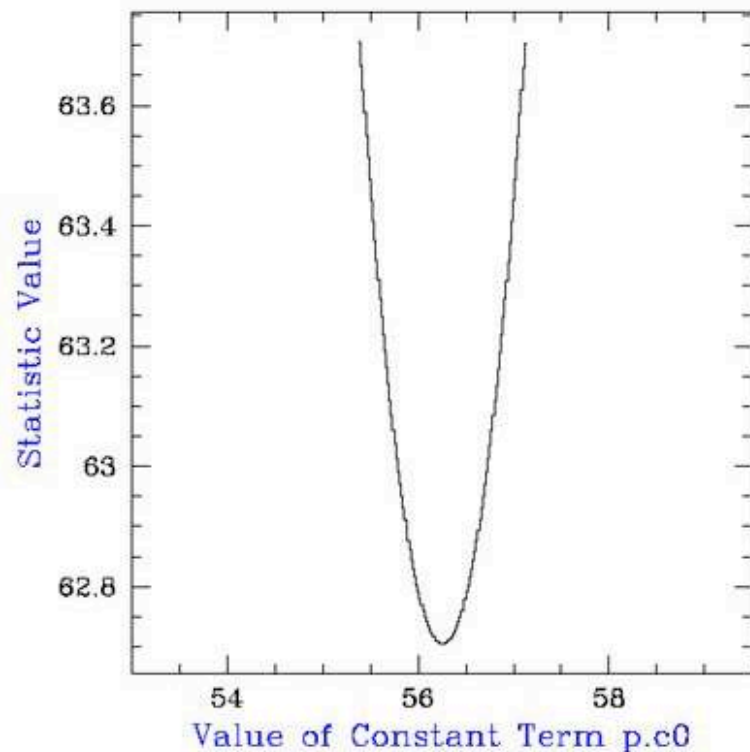
Calculating Confidence Limits means Exploring the Parameter Space – Statistical Surface



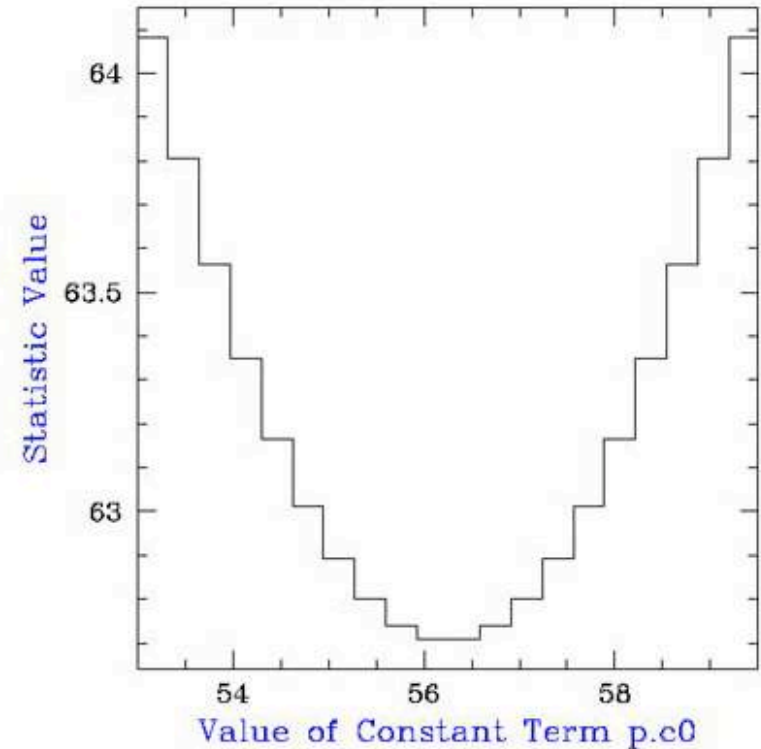
Example of a “well-behaved” statistical surface in parameter space, viewed as a multi-dimensional paraboloid (χ^2 , *top*), and as a multi-dimensional Gaussian ($\exp(-\chi^2/2) \approx L$, *bottom*).

Behaviour of Statistics for One Parameter

Interval – Uncertainty



Interval – Projection



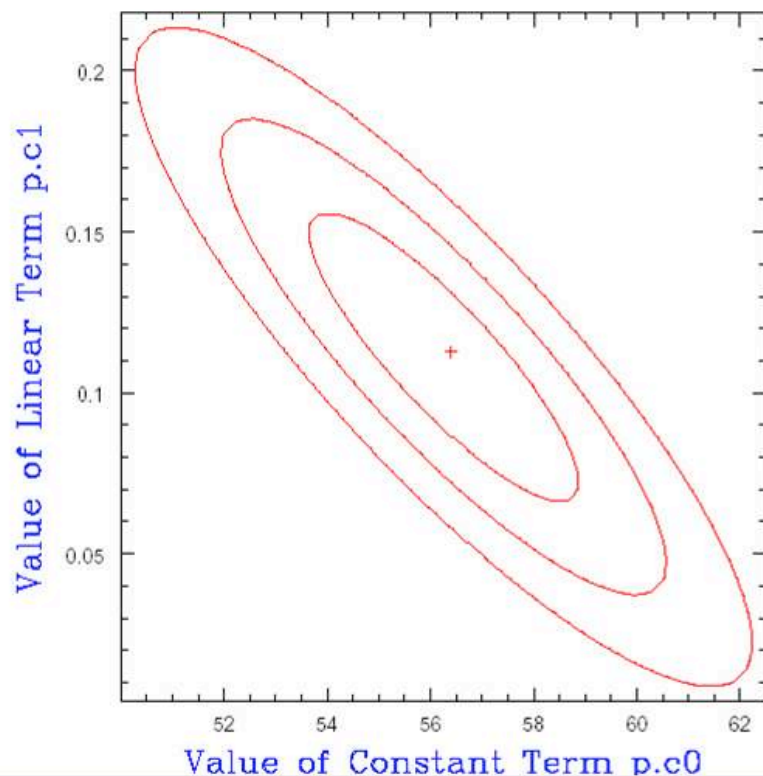
Comparison of Two methods in Sherpa

Confidence Limits for Two Parameters

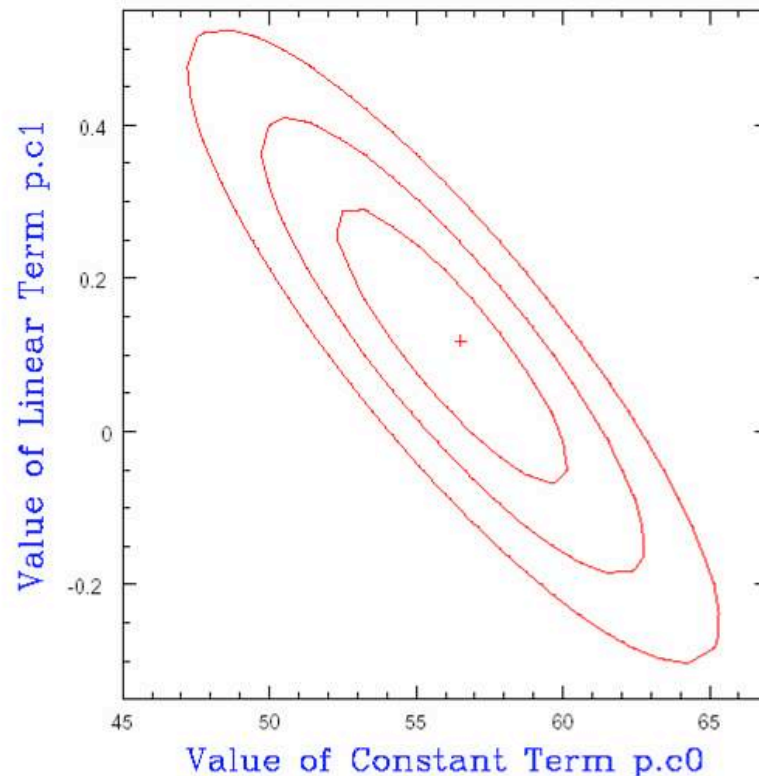
+ *Best fit parameters*

1 σ , 2 σ , 3 σ contours

Confidence Region – Uncertainty



Confidence Region – Projection



Comparison of Two methods in Sherpa

“Versions” of the χ^2 Statistic

The version of χ^2 derived above is dubbed “data variance” χ^2 , or χ_d^2 , because of the presence of D in the denominator. Generally, the χ^2 statistic is written as:

$$\chi^2 \equiv \sum_i^N \frac{(D_i - M_i)^2}{\sigma_i^2},$$

where σ_i^2 represents the (unknown!) variance of the Poisson distribution from which D_i is sampled.

χ^2 Statistic	σ_i^2
Data Variance	D_i
Model Variance	M_i
Gehrels	$[1 + \sqrt{D_i + 0.75}]^2$
Primini	M_i from previous best-fit
Churazov	based on <i>smoothed</i> data D
“Parent”	$\frac{\sum_{i=1}^N D_i}{N}$
Least Squares	1

Note that some X-ray data analysis routines may estimate σ_i for you during data reduction. In PHA files, such estimates are recorded in the **STAT_ERR** column.

Statistical Issues

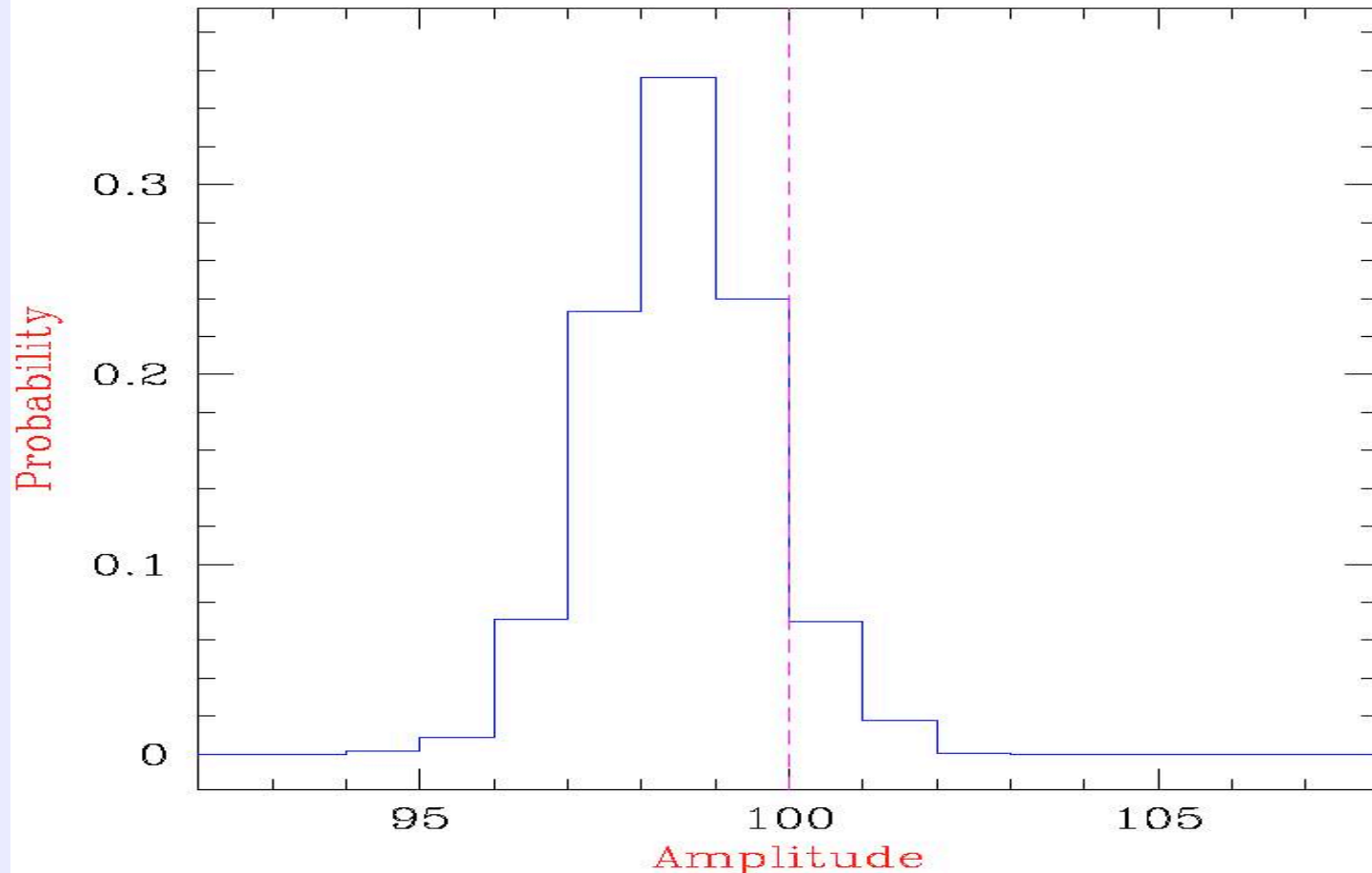
- Bias
- Goodness of Fit
- Background Subtraction
- Rebinning
- Errors

Statistical Issues: Bias

- If one samples a large number of datasets from a given model $M(\hat{\theta})$ and then fits this same model to these datasets (while letting θ vary), one will build up sampling distributions for each parameter θ_k .
- An estimator (**e.g.**, χ^2) is biased if the mean of these distributions ($E[\theta_k]$) differs from the true values $\theta_{k,o}$.
- The Poisson likelihood is an unbiased estimator.
- The χ^2 statistic **can** be biased, depending upon the choice of σ :
 - Using the *Sherpa* utility **FAKEIT**, we simulated 500 datasets from a constant model with amplitude 100 counts.
 - We then fit each dataset with a constant model, recording the inferred amplitude.

Statistic	Mean Amplitude
Gehrels	99.05
Data Variance	99.02
Model Variance	100.47
“Parent”	99.94
Primini	99.94
Cash	99.98

χ^2 Data Variance – Bias



A demonstration of bias. Five hundred datasets are sampled from a constant model with amplitude 100 and then are fit with the same constant amplitude model, using χ^2 with data variance. The mean of the distribution of fit amplitude values is not 100, as it would be if the statistic were an unbiased estimator.

Statistical Issues: Goodness-of-Fit

- The χ^2 goodness-of-fit is derived by computing

$$\begin{aligned}\alpha_{\chi^2} &= \int_{\chi_{\text{obs}}^2}^{\infty} d\chi^2 p(\chi^2 | N - P) \\ &= \frac{1}{2\Gamma\left(\frac{N-P}{2}\right)} \int_{\chi_{\text{obs}}^2}^{\infty} d\chi^2 \left(\frac{\chi^2}{2}\right)^{\frac{N-P}{2}-1} e^{-\frac{\chi^2}{2}}.\end{aligned}$$

This can be computed numerically using, *e.g.*, the **GAMMQ** routine of *Numerical Recipes*.

- A typical criterion for rejecting a model is $\alpha_{\chi^2} < 0.05$ (the “95% criterion”). However, using this criterion blindly ***is not recommended!***
- A quick’n’dirty approach to building intuition about how well your model fits the data is to use the **reduced** χ^2 , *i.e.*,

$$\chi_{\text{obs,r}}^2 = \chi_{\text{obs}}^2 / (N - P) :$$

- A “good” fit has $\chi_{\text{obs,r}}^2 \approx 1$.
- If $\chi_{\text{obs,r}}^2 \rightarrow 0$ the fit is “too good” -- which means (1) the errorbars are too large, (2) χ_{obs}^2 is **not** sampled from the χ^2 distribution, and/or (3) the data have been fudged.

The reduced χ^2 should never be used in any mathematical computation if you are using it, you are probably doing something wrong!

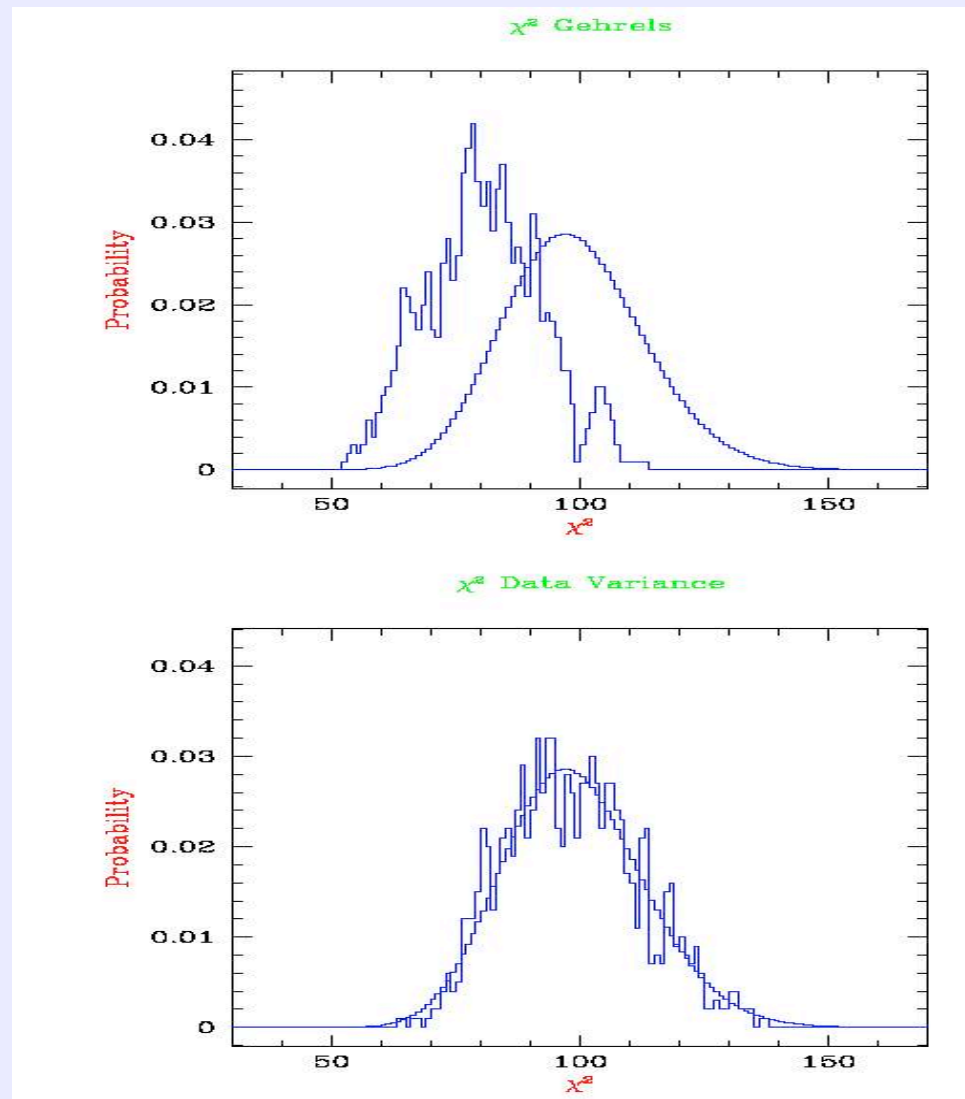


Figure 7: Comparison of the distributions of 500 sampled values of χ^2 versus the expected distribution for 99 degrees of freedom. Top: χ^2 with Gehrels variance. Bottom: χ^2 with data variance.

Statistical Issues: Background Subtraction

- A typical “dataset” may contain multiple spectra, one containing source and “background” counts, and one or more others containing only “background” counts.
 - The “background” may contain cosmic and particle contributions, *etc.*, but we'll ignore this complication and drop the quote marks.
- If possible, one should model background data:
 - ⇒ Simultaneously fit a background model M_B to the background dataset(s) B_j , and a source plus back-ground model $M_S + M_B$ to the raw dataset D .
 - ⇒ The background model parameters must have the same values in both fits, *i.e.*, do not fit the background data first, separately.
 - ⇒ Maximize $L_b \times L_{S+B}$ or minimize $\chi_B^2 + \chi_{S+B}^2$.
- However, many X-ray astronomers continue to subtract the background data from the raw data:

$$D'_i = D_i - \beta_D t_D \left[\frac{\sum_{j=1}^n B_{i,j}}{\sum_{j=1}^n \beta_{B_j} t_{B_j}} \right].$$

n is the number of background datasets, t is the observation time, and β is the “backscale” (given by the BACKSCAL header keyword value in a PHA file), typically defined as the ratio of data extraction area to total detector area.

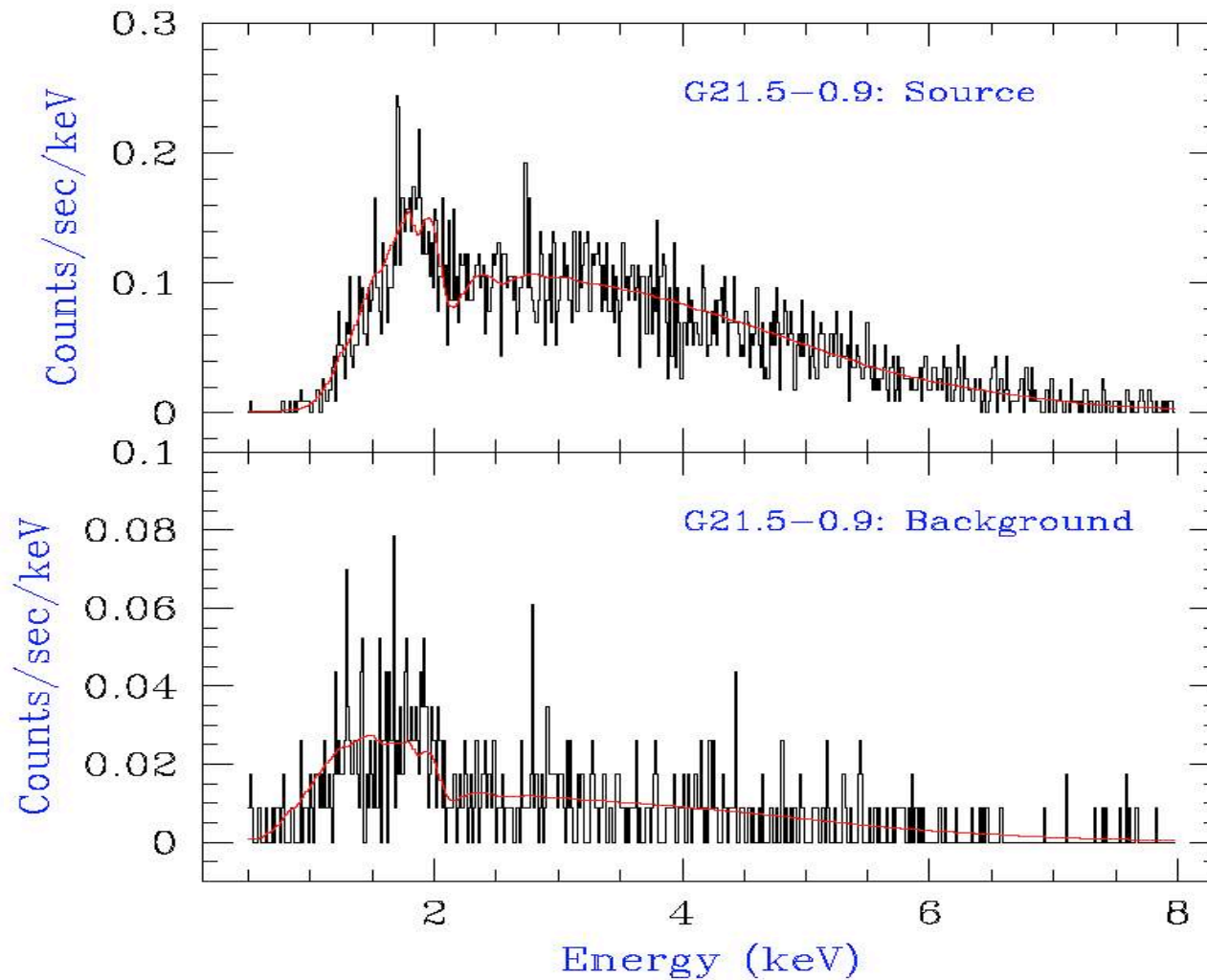


Figure 8: *Top*: Best-fit of a power-law times galactic absorption model to the source spectrum of supernova remnant G21.5-0.9. *Bottom*: Best-fit of a separate power-law times galactic absorption model to the background spectrum extracted for the same source.

Statistical Issues: Background Subtraction

- Why subtract the background?
 - It may be difficult to select an appropriate model shape for the background.
 - Analysis proceeds faster, since background datasets are not fit.
 - “It won't make any difference to the final results.”
- Why not subtract the background?

- The data D_i are not Poisson-distributed -- one cannot fit them with the Poisson likelihood. (Variances are estimated via **error propagation**:

$$\begin{aligned}V[f\{X_1, \dots, X_m\}] &\approx \sum_{i=1}^m \sum_{j=1}^m \frac{\partial f}{\partial \mu_i} \frac{\partial f}{\partial \mu_j} \text{cov}(X_i, X_j) \\ &\approx \sum_{i=1}^m \left(\frac{\partial f}{\partial \mu_i} \right)^2 V[X_i] \\ \Rightarrow V[D_i] &\approx V[D_i] + \sum_{j=1}^n \left(\frac{\beta_D t_D}{\beta_{B_j} t_{B_j}} \right)^2 V[B_{i,j}].\end{aligned}$$

- It may well make a difference to the final results:
 - * Subtraction reduces the amount of statistical information in the analysis quantitative accuracy is thus reduced.
 - * Fluctuations can have an adverse effect, in, **e.g., line detection.**

Statistical Issues: Rebinning

- **Rebinning data invariably leads to a loss of statistical information!**
- Rebinning is not necessary if one uses the Poisson likelihood to make statistical inferences.
- However, the rebinning of data may be necessary to use χ^2 statistics, if the number of counts in any bin is ≤ 5 . In X-ray astronomy, rebinning (or **grouping**) of data may be accomplished with:
 - **grppha**, an **FTOOLS** routine; or
 - **dmgroup**, a **CIAO** Data Model Library routine.

One common criterion is to sum the data in adjacent bins until the sum equals five (or more).

- **Caveat:** always estimate the errors in rebinned spectra using the new data D in each new bin (since these data are still Poisson-distributed), rather than propagating the errors in each old bin.
 - ⇒ For example, if three bins with numbers of counts 1, 3, and 1 are grouped to make one bin with 5 counts, one should estimate $V[D' = 5]$ and *not* $V[D'] = V[D_1 = 1] + V[D_2 = 3] + V[D_3 = 1]$. The propagated errors may overestimate the true errors.

Statistical Issues: Systematic Errors

- In X-ray astronomy, one usually speaks of two types of errors: statistical errors, and systematic errors.
- Systematic errors are uncertainties in instrumental calibration. For instance:
 - Assume a spectrum observed for time t with a telescope with perfect resolution and an effective area A_i . Furthermore, assume that the uncertainty in A_i is $\sigma_{A,i}$.
 - Neglecting data sampling, in bin i , the expected number of counts is $D_i = D_{\gamma,i}(\Delta E)tA_i$.
 - We estimate the uncertainty in D_i as
$$\sigma_{D_i} = D_{\gamma,i}(\Delta E)t\sigma_{A,i} = D_{\gamma,i}(\Delta E)tf_iA_i = f_iD_i$$
- The systematic error f_iD_i ; in PHA files, the quantity f_i is recorded in the SYS_ERR column.
- Systematic errors are added in quadrature with statistical errors; for instance, if one uses χ^2_d to assess the quality of fit, then $\sigma_i = \sqrt{D_i + (f_iD_i)^2}$.
- To use information about systematic errors in a Poisson likelihood fit, one must incorporate this information into the model, as opposed to simply adjusting the estimated error for each datum.

Summary

- Motivation: why do we need statistics?
- Probabilities/Distributions
- Poisson Likelihood
- Parameter Estimation
- Statistical Issues
- Statistical Tests – still to come....

Conclusions

Statistics is the main tool for any astronomer who need to do data analysis and need to decide about the physics presented in the observations.

References:

Peter Freeman's Lectures from the Past X-ray Astronomy School:
<http://xrayschool.gsfc.nasa.gov/docs/xrayschool-2003/talks.html>

“Practical Statistics for Astronomers”, Wall & Jenkins, 2003
Cambridge University Press

Eadie et al 1976, “Statistical Methods in Experimental Physics”

Selected References

● General statistics:

- Babu, G. J., Feigelson, E. D. 1996, *Astrostatistics* (London: Chapman & Hall)
- Eadie, W. T., Drijard, D., James, F. E., Roos, M., & Sadoulet, B. 1971, *Statistical Methods in Experimental Physics* (Amsterdam: North-Holland)
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, *Numerical Recipes* (Cambridge: Cambridge Univ. Press)

● Introduction to Bayesian Statistics:

- Lored, T. J. 1992, in *Statistical Challenges in Modern Astronomy*, ed. E. Feigelson & G. Babu (New York: Springer-Verlag), 275

● Modified \mathcal{L} and χ^2 Statistics:

- Cash, W. 1979, *ApJ* 228, 939
- Churazov, E., et al. 1996, *ApJ* 471, 673
- Gehrels, N. 1986, *ApJ* 303, 336
- Kearns, K., Primi, F., & Alexander, D. 1995, in *Astronomical Data Analysis Software and Systems IV*, eds. R. A. Shaw, H. E. Payne, & J. J. E. Hayes (San Francisco: ASP), 331

● Issues in Fitting:

- Freeman, P. E., et al. 1999, *ApJ* 524, 753 (and references therein)

● *Sherpa* and *XSPEC*:

- Freeman, P. E., Doe, S., & Siemiginowska, A. 2001, astro-ph/0108426
- http://asc.harvard.edu/ciao/download/doc/sherpa_html_manual/index.html
- Arnaud, K. A. 1996, in *Astronomical Data Analysis Software and Systems V*, eds. G. H. Jacoby & J. Barnes (San Francisco: ASP), 17
- <http://heasarc.gsfc.nasa.gov/docs/xanadu/xspec/manual/manual.html>

Properties of Distributions

The beginning X-ray astronomer only needs to be familiar with four properties of distributions: the mean, mode, variance, and standard deviation, or “error.”

- Mean: $\mu = E[X] = \int dX X p(X)$
- Mode: $\max[p(X)]$
- Variance: $V[X] = E[(X - \mu)^2] = \int dX (X - \mu)^2 p(X)$
- Error: $\sigma_x = \sqrt{V[X]}$

Note that if the distribution is Gaussian, then σ is indeed the Gaussian σ (hence the notation).

If two random variables are to be jointly considered, then the sampling distribution is two-dimensional, with shape locally described by the **covariance matrix**:

$$\begin{pmatrix} V[X_1] & \text{cov}[X_1, X_2] \\ \text{cov}[X_1, X_2] & V[X_2] \end{pmatrix} \quad \text{where} \quad \begin{aligned} \text{cov}[X_1, X_2] &= E[(X_1 - \mu_{x_1})(X_2 - \mu_{x_2})] \\ &= E[X_1 X_2] - E[X_1]E[X_2] \end{aligned}$$

The related **correlation coefficient** is $\text{corr}[X_1, X_2] = \frac{\text{cov}[X_1, X_2]}{\sigma_{x_1} \sigma_{x_2}}$.

The correlation coefficient can range from -1 to 1.