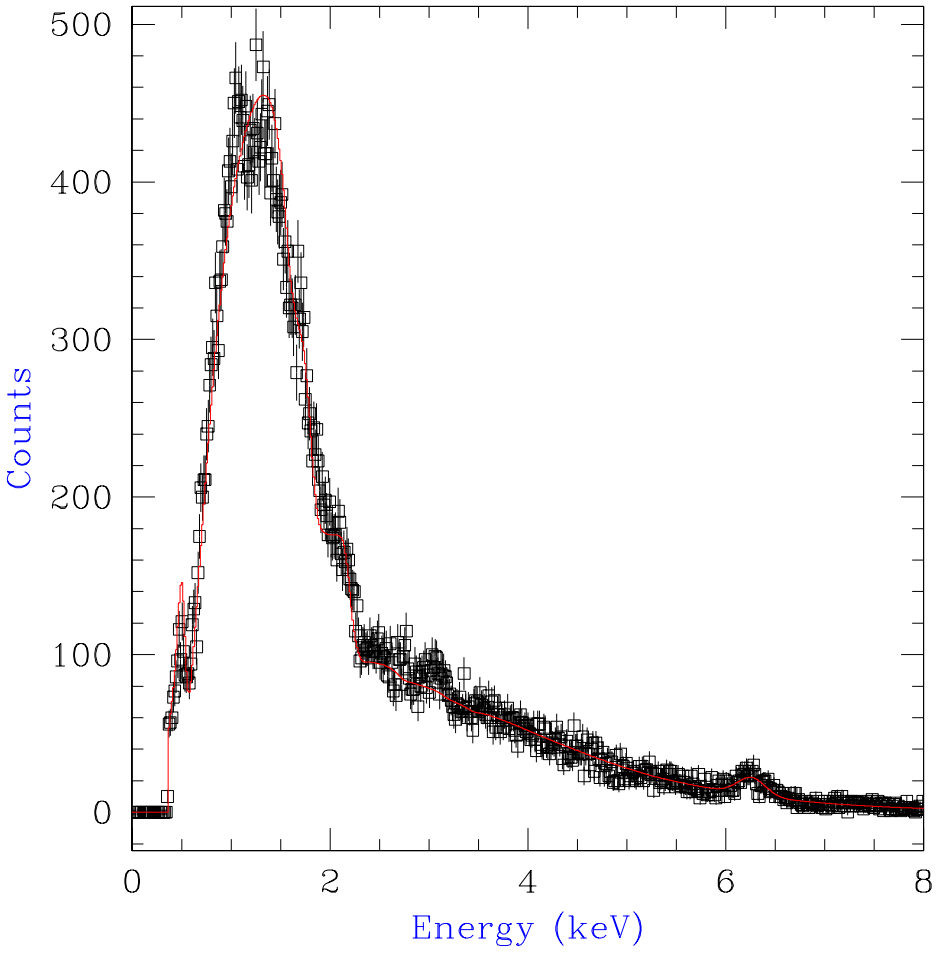
Statistics I:

Issues in Model Fitting

in the X-Ray Regime



Peter Freeman

Harvard-Smithsonian Center for Astrophysics

Glossary of Important Notation

- D: a dataset
- D_i: the datum of bin i of the dataset
- N: the number of bins in the dataset
- B: a background dataset associated with D
- B_i: the datum of bin i of the background dataset
- $M = M(\theta)$: a model with free parameters θ
- $\hat{\theta}$: the vector of *best-fit* model parameters
- P: the number of (freely varying) model parameters
- M_i: the *convolved* model amplitude in bin i
- μ: the mean of a distribution
- V: the variance of a distribution
- σ: the standard deviation of a distribution
- $E[X]$: the expectation of variable X
- $\mathcal{L}$: the likelihood
- L: the log-likelihood $\log\mathcal{L}$
- χ^2: the "chi-square" statistic

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 $(\mu = [\Sigma_{i=1}^{N} D_i]/N)$.
- *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 $p(X)$. Note that in the special case of the Gaussian (or Normal) distribution, $p(X)$ may be written as $N(\mu, \sigma^2)$, where μ is the Gaussian mean, and σ^2 is its variance.

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*:

$$\begin{pmatrix} V[X_1] & \mathrm{cov}[X_1, X_2] \\ \mathrm{cov}[X_1, X_2] & V[X_2] \end{pmatrix}$$

where

$$\begin{aligned} \mathrm{cov}[X_1, X_2] &= E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= E[X_1 X_2] - E[X_1] E[X_2] \end{aligned}$$

The related *correlation coefficient* is

$$\mathrm{corr}[X_1, X_2] = \frac{\mathrm{cov}[X_1, X_2]}{\sigma_{X_1} \sigma_{X_2}}.$$

The correlation coefficient can range from -1 to 1.

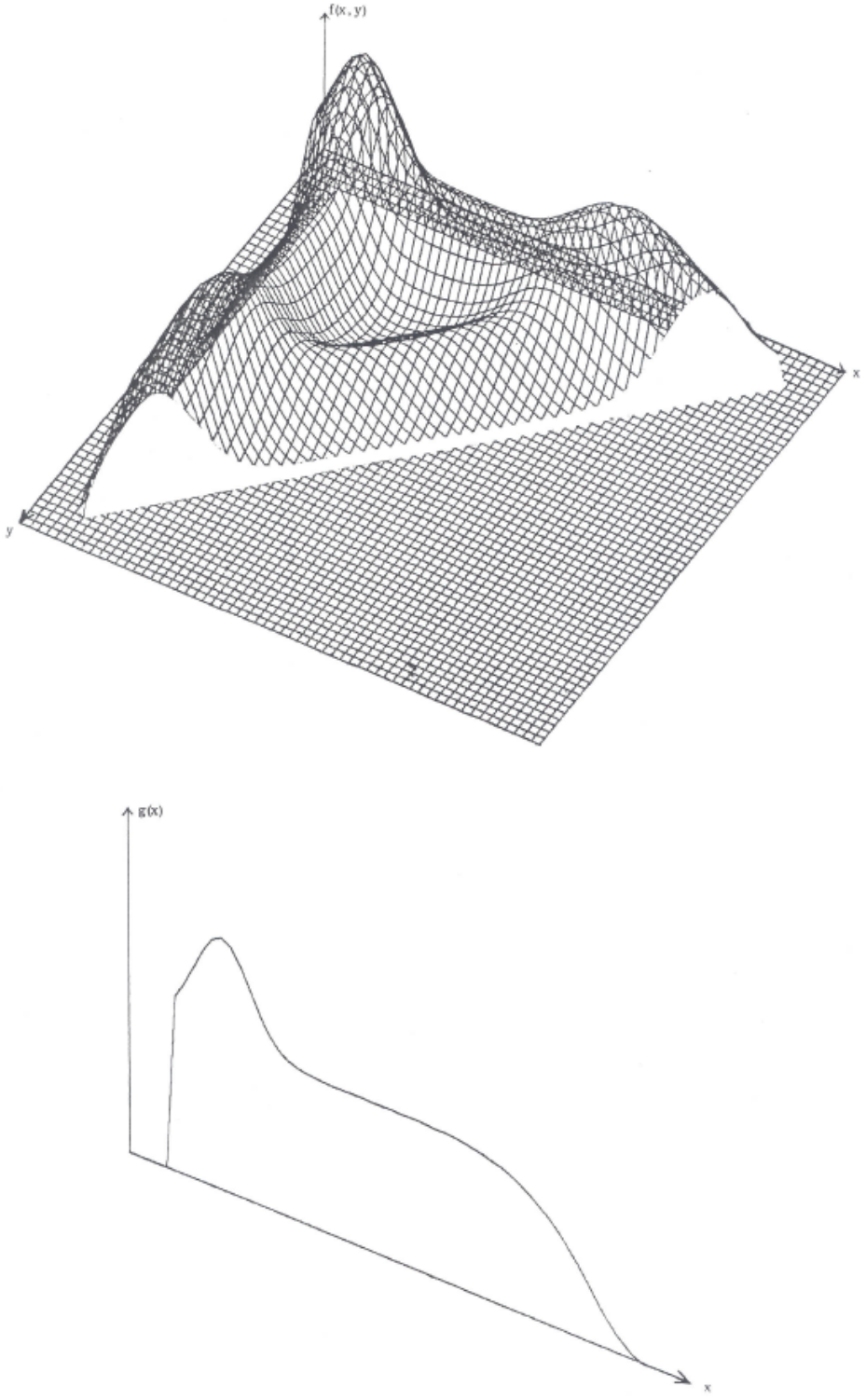


Figure 1: *Top:* example of a joint probability sampling distribution for two random variables. *Bottom:* the marginal sampling distribution $p(x) = \int dy p(x, y)$ (Eadie *et al.* 1971, p. 16).

The Poisson Distribution

In the remainder of this class, we will concentrate exclusively upon fitting counts spectra, i.e. fitting data sampled from the Poisson distribution.

The discrete Poisson distribution

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

gives the probability of finding exactly D_i events in bin i of dataset D in a given length of time, if the events occur independently at a constant rate M_i.

Things to remember about the Poisson distribution:

- $\mu = E[D_i] = M_i$;
- $V[D_i] = M_i$;
- $\mathrm{cov}[D_{i_1}, D_{i_2}] = 0$;
- the sum of n Poisson-distributed variables (found by, *e.g.* combining the data in n bins) is itself Poisson-distributed with variance $\Sigma_{i=1}^{n} M_i$; and
- as $M_i \to \infty$, *the Poisson distribution converges to a Gaussian distribution* $N(\mu = M_i, \sigma^2 = M_i)$.

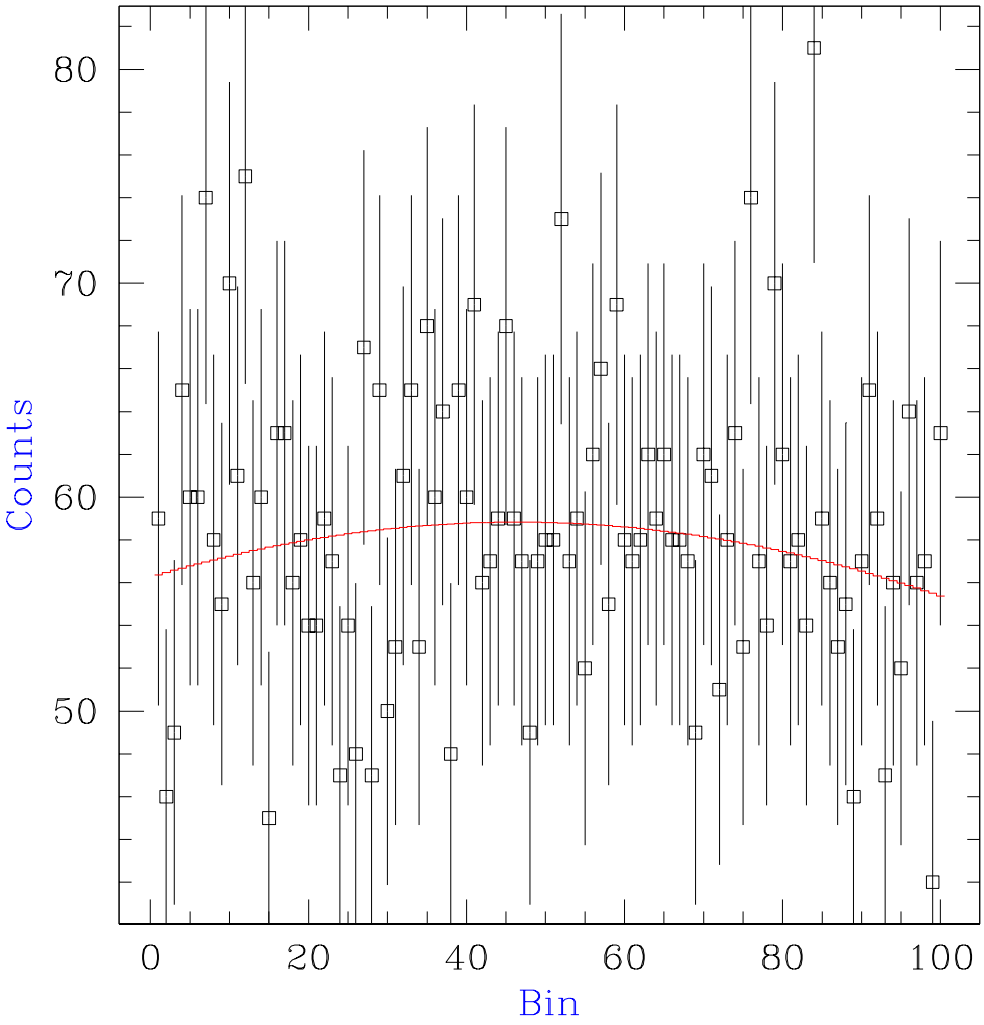


Figure 2: Integer counts spectrum sampled from a constant amplitude model with mean μ = 60 counts, and fit with a parabolic model.

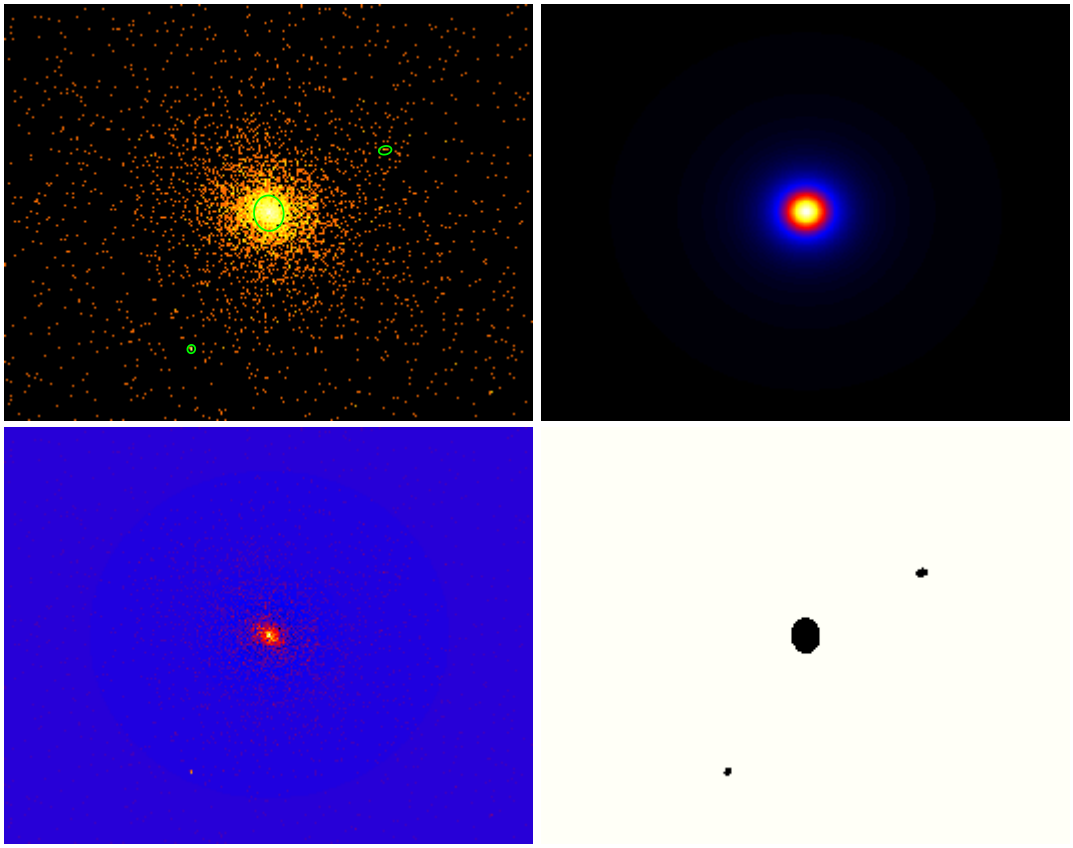


Figure 3: Example of a two-dimensional integer counts spectrum. *Top Left: Chandra* ACIS-S data of X-ray cluster MS 2137.3-2353, with *SAODS9* source regions superimposed. *Top Right:* Best-fit of a two-dimensional beta model to the filtered data. *Bottom Left:* Residuals (in units of σ) of the best fit. *Bottom Right:* The applied filter; the data within the ovals were excluded from the fit.

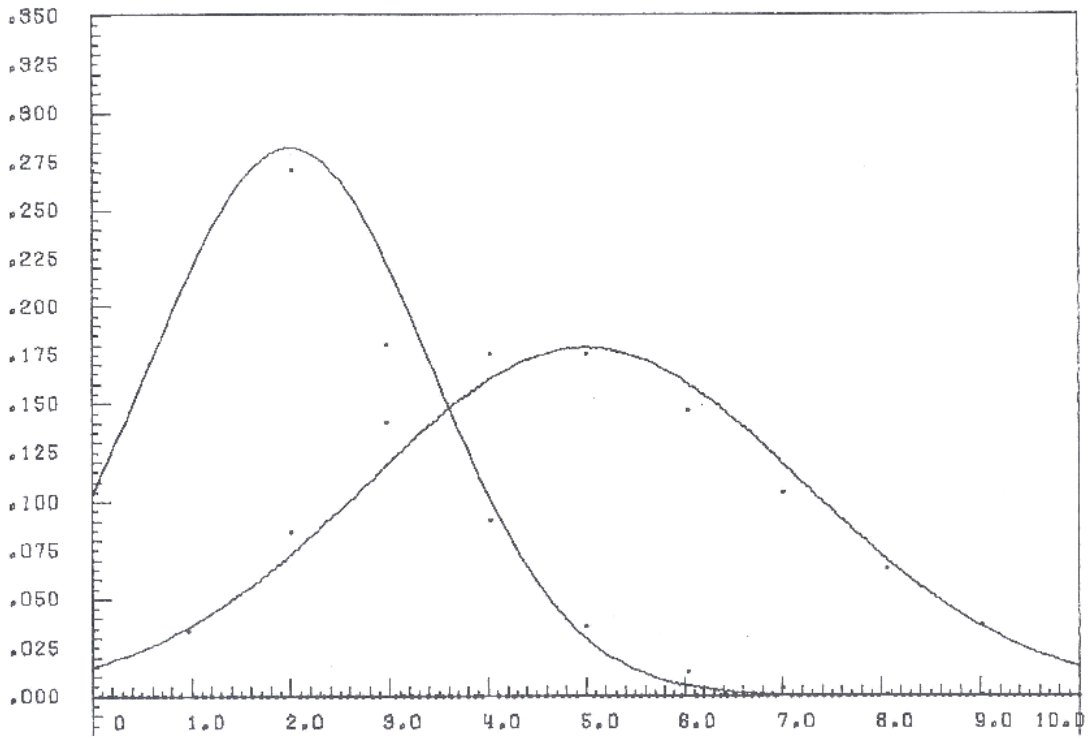


Figure 4: Comparison of Poisson distributions (dotted) of mean $\mu = 2$ and 5 with normal distributions of the same mean and variance (Eadie *et al.* 1971, p. 50).

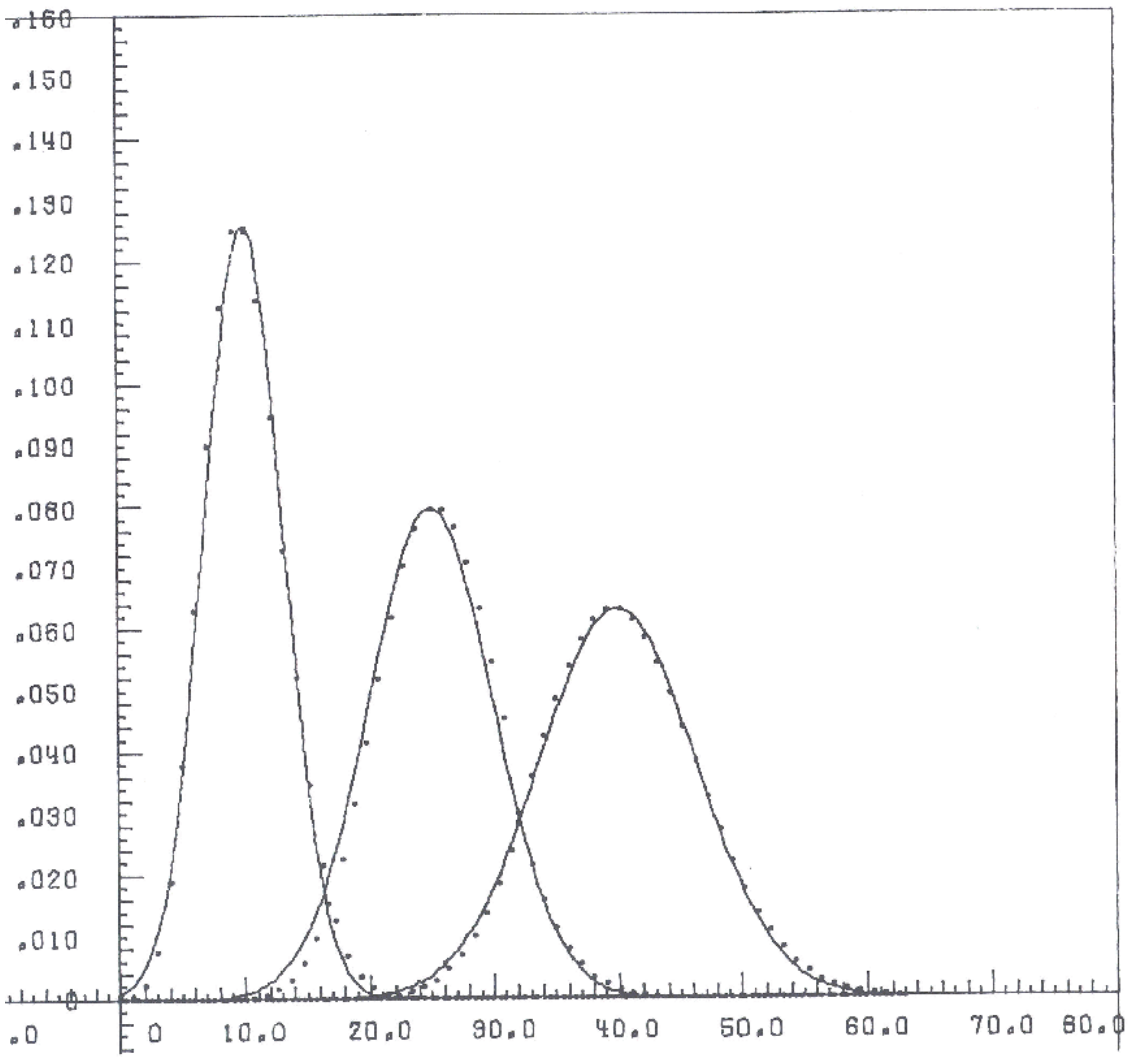


Figure 5: Comparison of Poisson distributions (dotted) of mean $\mu = 10$, 25 and 40 with normal distributions of the same mean and variance (Eadie *et al.* 1971, p. 50).

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*:

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

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 of *XSPEC*); and
- the fact that there is an alternative in the Gaussian limit: the χ^2 statistic.

The χ^2 Statistic

Here, we demonstrate the connection between the Poisson likelihood and the χ^2 statistic.[1]

- Step 1: write down the Poisson likelihood (in one bin).

$$\mathcal{L}_i = \frac{M_i^D}{D_i!}\exp(-M_i)$$

- Step 2: apply Stirling's approximation.

$$\begin{aligned} D_i! &= \sqrt{2\pi D_i} D_i^{D_i} e^{-D_i} \\ \Rightarrow \quad \mathcal{L}_i &\approx \frac{1}{\sqrt{2\pi D_i}} \left(\frac{M_i}{D_i}\right)^{D_i} e^{D_i - M_i} \end{aligned}$$

- Step 3: look near, *e.g.*, the log-likelihood peak, and reparameterize in terms of $\epsilon \equiv \frac{M_i - D_i}{\sqrt{D_i}}$.

$$\begin{aligned} L_i = \log\mathcal{L}_i &\approx -\frac{1}{2}\log(2\pi D_i) + D_i \log(\frac{M_i}{D_i}) + D_i - M_i \\ &\approx -\frac{1}{2}\log(2\pi D_i) + D_i \log(1 + \frac{\epsilon}{\sqrt{D_i}}) - \epsilon\sqrt{D_i} \\ &\approx -\frac{1}{2}\log(2\pi D_i) + \\ &\qquad D_i\left(\frac{\epsilon}{\sqrt{D_i}} - \frac{\epsilon^2}{2D_i} + \frac{\epsilon^3}{3D_i^{3/2}} - \cdots\right) - \epsilon\sqrt{D_i} \\ &\approx -\frac{1}{2}\log(2\pi D_i) - \frac{\epsilon^2}{2} + \mathcal{O}(\frac{\epsilon^3}{\sqrt{D_i}}) \end{aligned}$$

$$\Rightarrow \quad \mathcal{L}_i \approx \frac{1}{\sqrt{2\pi D_i}} \exp\left[-\frac{(M_i - D_i)^2}{2D_i}\right] \propto \exp\left(-\frac{\chi^2}{2}\right)$$

[1] The following is based on unpublished notes by Loredo (1993).

Validity of the χ^2 Statistic

Summarizing the results shown on the last panel, if

- $D_i \gg 1$ in every bin i, and
- terms of order ϵ^3 and higher in the Taylor series expansion of L may be ignored,

then the statistic χ^2 may be used to estimate the Poisson likelihood, and an observed value $\chi^2_{\rm obs}$ will be sampled from the χ^2 distribution for $N - P$ degrees of freedom.

$\Rightarrow$ Regarding the first condition above, the general rule-of-thumb is that there should be a minimum of five counts in every bin.

$\Rightarrow$ Regarding the second condition above, it is only an major issue if the fit is *bad*.

- However, bad fits are common in X-ray astronomy; one example is the fit of a continuum model to data exhibiting an obvious (emission or absorption) line. Inferences made using such a fit *can be suspect!*

Note that if either rule breaks down, you can still use the χ^2 statistic; however, it will no longer be χ^2-distributed and you may need to use Monte Carlo simulations to make statistical inferences. Also, your estimates of best-fit parameter values may not closely match estimates you would have made using the Poisson likelihood.

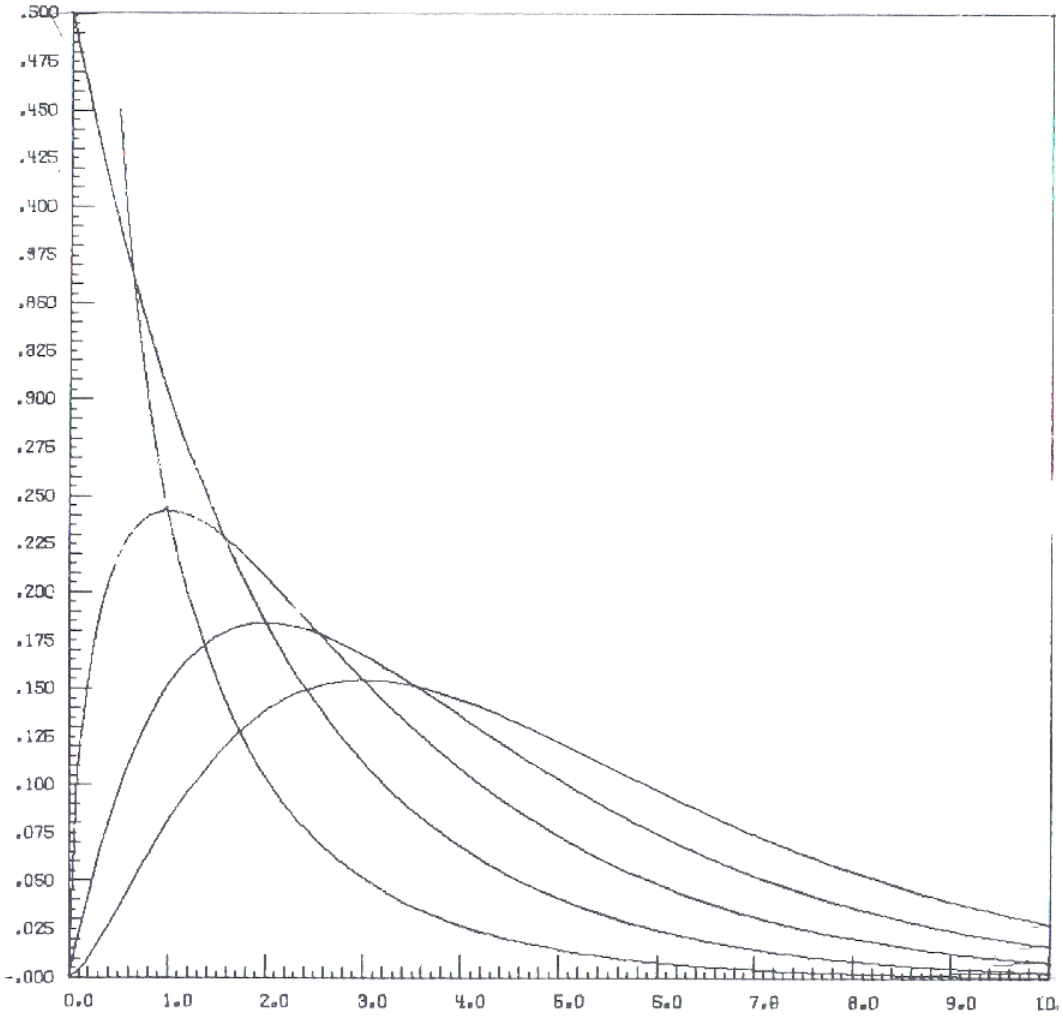


Figure 6: Examples of the χ^2 distribution for $\nu = N - P = 1$, 2, 3, 4, and 5 (Eadie *et al.* 1971, p. 64).

Versions of the χ^2 Statistic

The version of χ^2 derived above is dubbed "data variance" χ^2, or χ^2_d, 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 *smoothed* 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: Goodness-of-Fit

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

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

  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.* $\chi^2_{\rm obs,r} = \chi^2_{\rm obs}/(N-P)$:
  - A "good" fit has $\chi^2_{\rm obs,r} \approx 1$.
  - If $\chi^2_{\rm obs,r} \to 0$, the fit is "too good"—which means (1) the errorbars are too large, (2) $\chi^2_{\rm obs}$ 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!

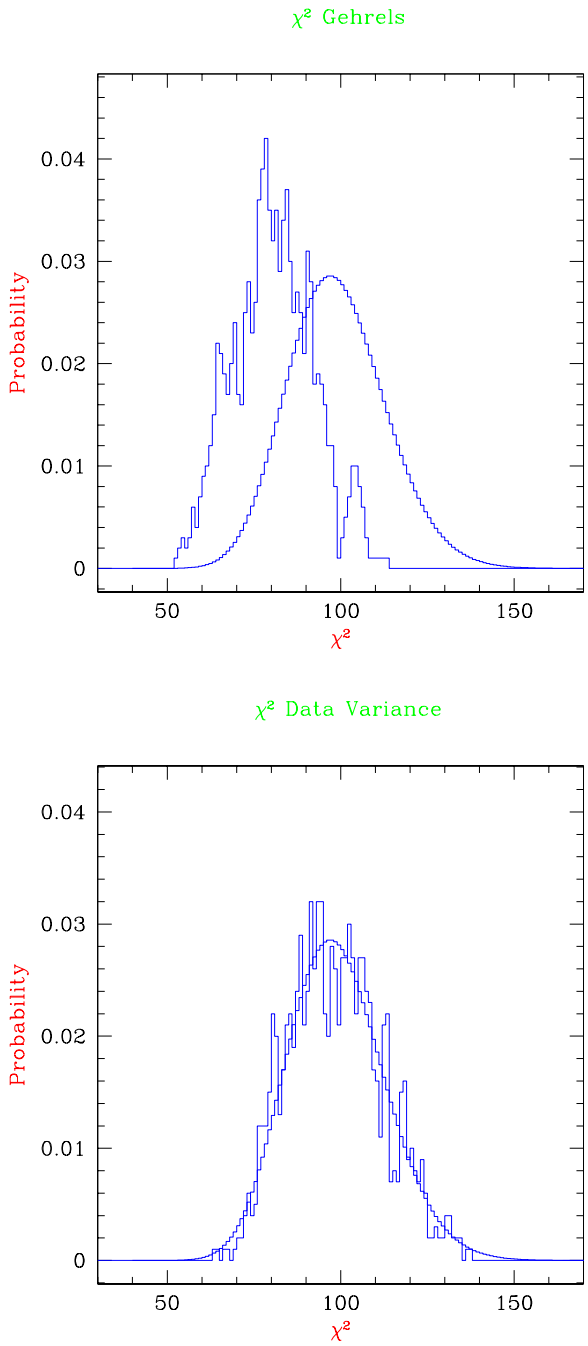


Figure 7: Comparison of the distributions of 500 sampled values of χ^2 versus the expected distribution for 99 degrees of freedom. *Top:* χ^2 with Gehrels variance. *Bottom:* χ^2 with data variance.

Statistical Issues: Background Subtraction

- A typical "dataset" may contain multiple spectra, one of which contains contributions from the source of interest and the background, and one or more others which contain background counts alone. (The background itself may contain contributions from the cosmic X-ray background, the particle background, *etc.*, but we'll ignore this complication.)
- The proper way to treat background data is to model them!

```
sherpa> data source.pi
sherpa> back back.pi
sherpa> source = xswabs[sabs]*pow[sp]
sherpa> bg = xswabs[babs]*pow[bp]
sherpa> statistic cash
sherpa> fit  # maximize L(B)*L(S+B) or minimize X^2(B)+X^2(S+B)
...
 powll:   final function value    =    -7.01632E+03
          sabs.nH  2.35843  10^22/cm^2
            sp.gamma  1.48526
            sp.ampl  0.00195891
          babs.nH  0.671569  10^22/cm^2
            bp.gamma  1.07225
            bp.ampl  0.000107204
sherpa> projection
...
       ------------------------------------------------------
       Parameter Name      Best-Fit Lower Bound     Upper Bound
       ------------------------------------------------------
         sabs.nH            2.35732  -0.0981442      +0.150539
           sp.gamma          1.48477  -0.0645673      +0.101794
           sp.ampl       0.00195682  -0.000177659    +0.000317947
```

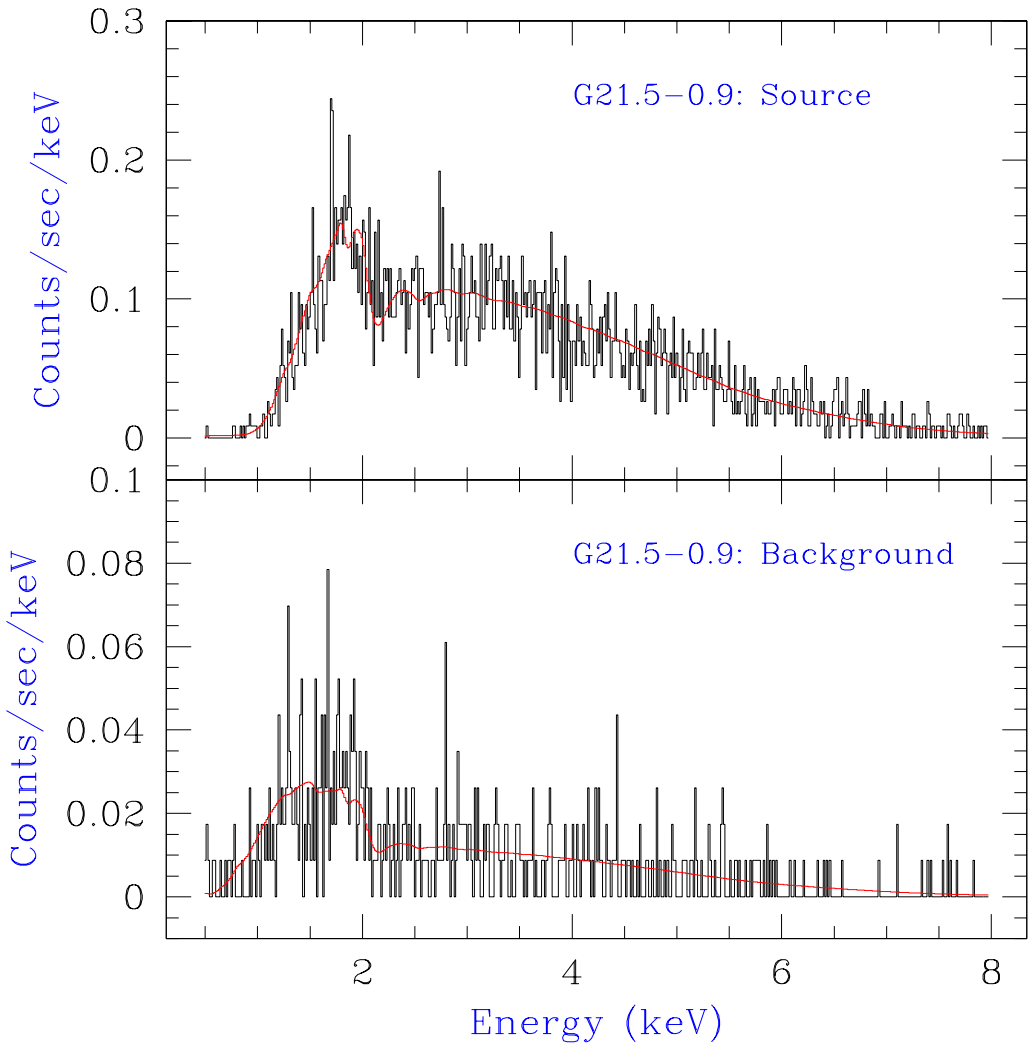


Figure 8: *Top:* Best-fit of a power-law times galactic absorption model to the source spectrum of supernova remnant G21.5-0.9. *Bottom:* Best-fit of a separate power-law times galactic absorption model to the background spectrum extracted for the same source.

Statistical Issues: Background Subtraction

- However, many X-ray astronomers *subtract* background data from the raw data:

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

  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).

- Why should one not subtract background?
  - It reduces the amount of statistical information in the analysis—the final fit parameter values will be a less accurate estimate of the true values.
  - The data D_i' are not Poisson-distributed—one cannot fit them with the Poisson likelihood (or the Cash statistic), even in the low-count limit.
  - Fluctuations (particularly in the vicinity of localized features) can adversely affect analysis.

- To use χ^2, the *errors must be propagated*:

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

Statistical Issues: Background Subtraction

- Here, we repeat the fit from above, except that this time the data are background-subtracted:

```
sherpa> data source.pi
sherpa> back back.pi
sherpa> subtract
sherpa> statistic chi gehrels  # can't use Cash!
sherpa> fit
...
 powll:   final function value    =     1.88299E+02
          sabs.nH  2.67251  10^22/cm^2
            sp.gamma  1.74921
            sp.ampl  0.00261343

sherpa> projection
...
Computed for projection.sigma = 1
        ------------------------------------------------------
        Parameter Name      Best-Fit Lower Bound     Upper Bound
        ------------------------------------------------------
          sabs.nH            2.67251  -0.202747       +0.214219
            sp.gamma         1.74921  -0.14036        +0.144823
            sp.ampl       0.00261343  -0.000475006    +0.000597735
```

- Compare this with the previous result:

```
        ------------------------------------------------------
        Parameter Name      Best-Fit Lower Bound     Upper Bound
        ------------------------------------------------------
          sabs.nH            2.35732  -0.0981442      +0.150539
            sp.gamma         1.48477  -0.0645673      +0.101794
            sp.ampl       0.00195682  -0.000177659    +0.000317947
```

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 $\lesssim 5$. 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'_i in each new bin (since these data are still Poisson-distributed), rather than propagating the errors in each old bin.

  $\Rightarrow$ 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: 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

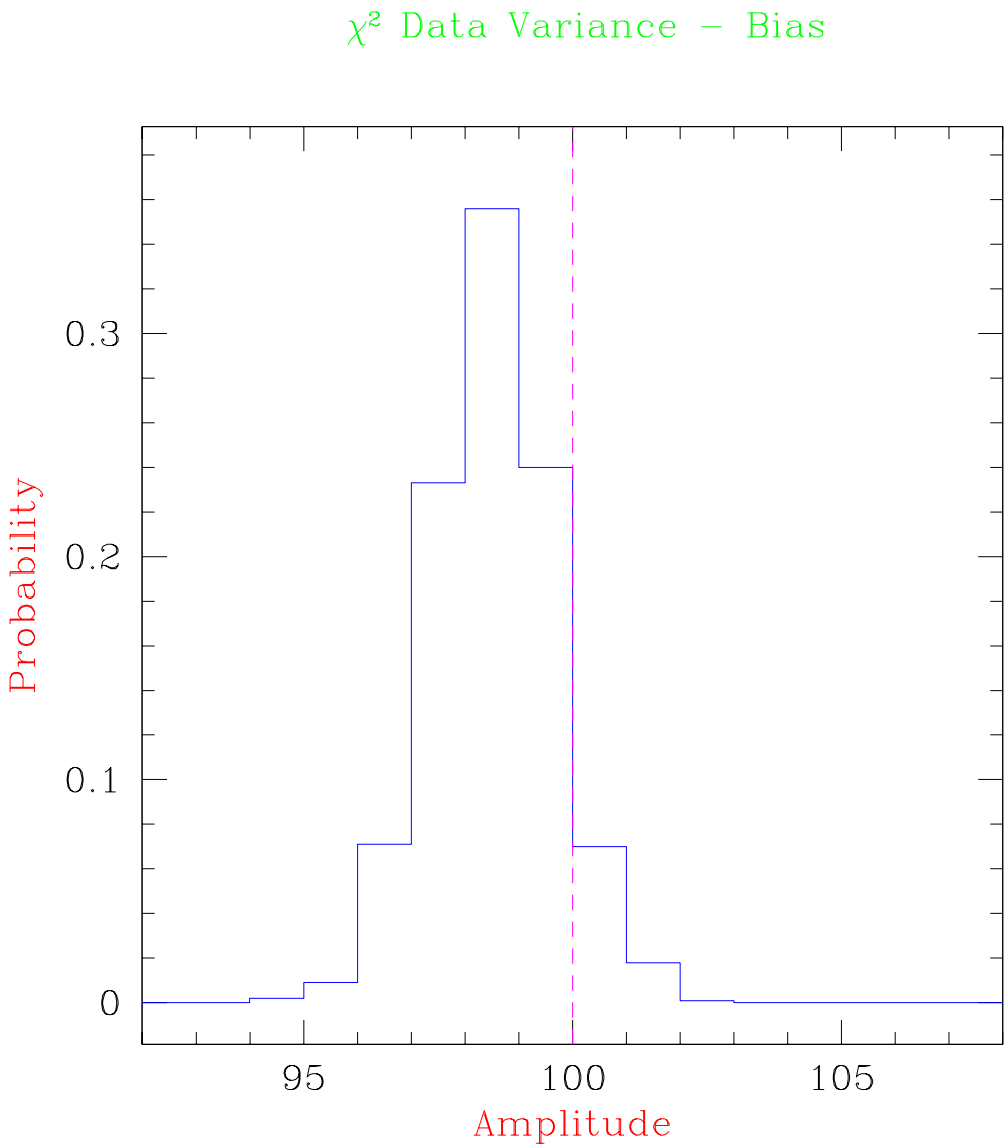


Figure 9: A demonstration of bias. Five hundred datasets are sampled from a constant model with amplitude 100 and then are fit with the same constant amplitude model, using χ^2 with data variance. The mean of the distribution of fit amplitude values is not 100, as it would be if the statistic were an unbiased estimator.

Statistical Issues: Systematic Errors

- In X-ray astronomy, one usually speaks of two types of errors: statistical errors, and systematic errors.
- Systematic errors reflect uncertainties in instrumental calibration. For instance:
  - Assume a flat 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.

Methodologies

It is important to note that the field of statistics may be roughly divided into two schools: the so-called "frequentist" (or classical) school, and the Bayesian school.

- A frequentist assesses a model $M(\hat{\theta})$ by first assuming that
  - M is the "true" model, and
  - $\hat{\theta}$ are the "true" model parameter values,

  and then comparing the probability of observing the dataset D with the probabilities of observing *other datasets predicted by* M.
- A Bayesian assesses $M(\hat{\theta})$ by comparing its probability (given the observed dataset D only) with the probabilities of *other parameterized models*, given D.

We have been able to ignore the differences between the two methodologies when discussing model fitting, up to now.

Statistical Issues: Bayesian Fitting

The centerpiece of the Bayesian statistical methodology is Bayes' theorem. As applied in a model fit, it may be written as

$$p(\theta|D) \;=\; p(\theta)\frac{p(D|\theta)}{p(D)}\,,$$

where

- $p(D|\theta)$ is the likelihood $\mathcal{L}$ (which may be estimated as $\exp(-\chi^2/2)$);
- $p(\theta)$ is the *prior distribution* for θ, reflecting your knowledge of the parameter values *before* the experiment;
- $p(\theta|D)$ is the *posterior distribution* for θ, reflecting your knowledge of the parameter values *after* the experiment; and
- $p(D)$ is an ignorable normalization constant.

For now, keep in mind that a Bayesian is more interested in finding the mode of the posterior distribution than in determining the maximum likelihood! (Delving into the hurly-burly world of prior specification is beyond the scope of this class...which is now over!)
