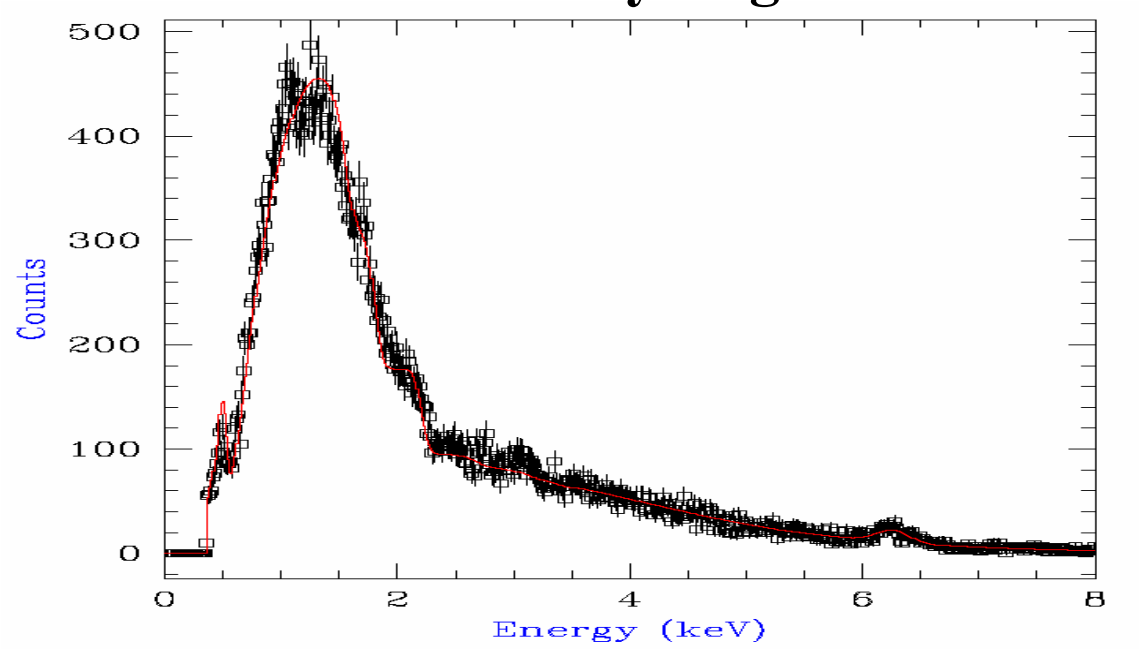
Statistics I:
Issues in Model Fitting
in the X-Ray Regime



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

- *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: $\mu = \mathrm{E}[X] = \int dX\, X\, p(X)$
- 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.

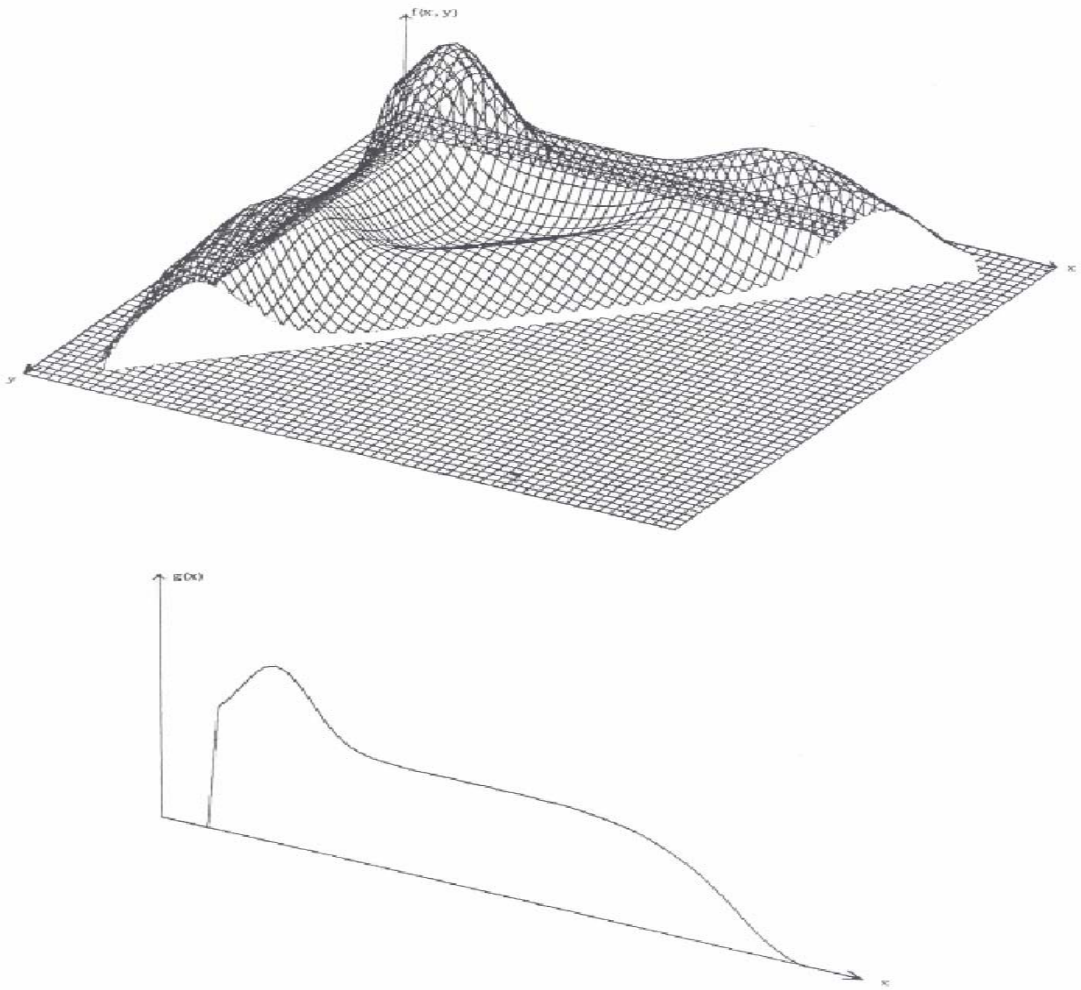


Figure 1: *Top:* example of a joint probability sampling distribution for two random variables. *Bottom:* the marginal sampling distribution $p(x) = \int dy\, p(x,y)$ (Eadie *et al*. 1971, p. 16).

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

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

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 $M_i \rightarrow \infty$, *the Poisson distribution converges to a Gaussian distribution* $N(\mu = M_i ; \sigma^2 = M_i)$.

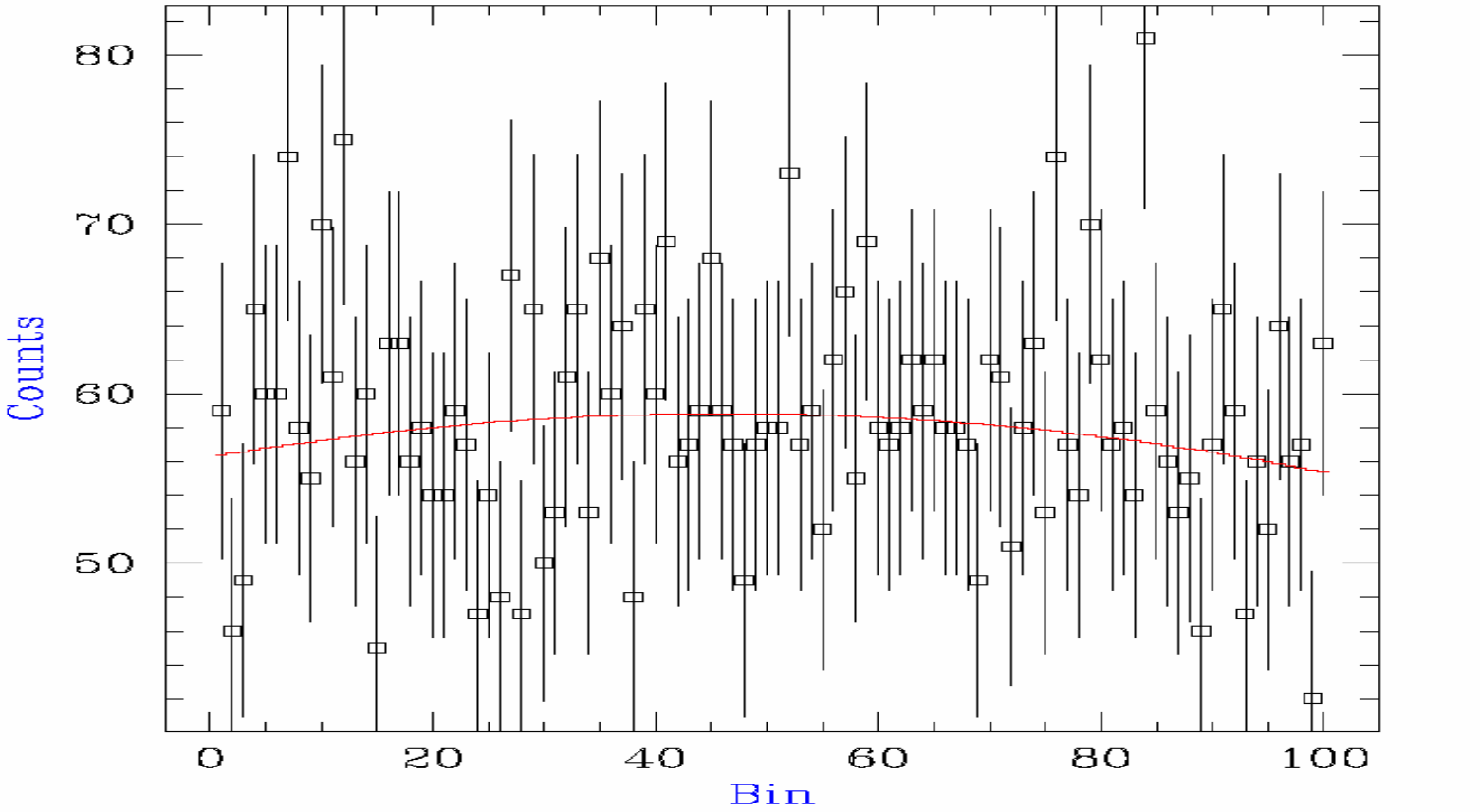


Figure 2: Integer counts spectrum sampled from a constant amplitude model with mean μ = 60 counts, and fit with a parabolic model.

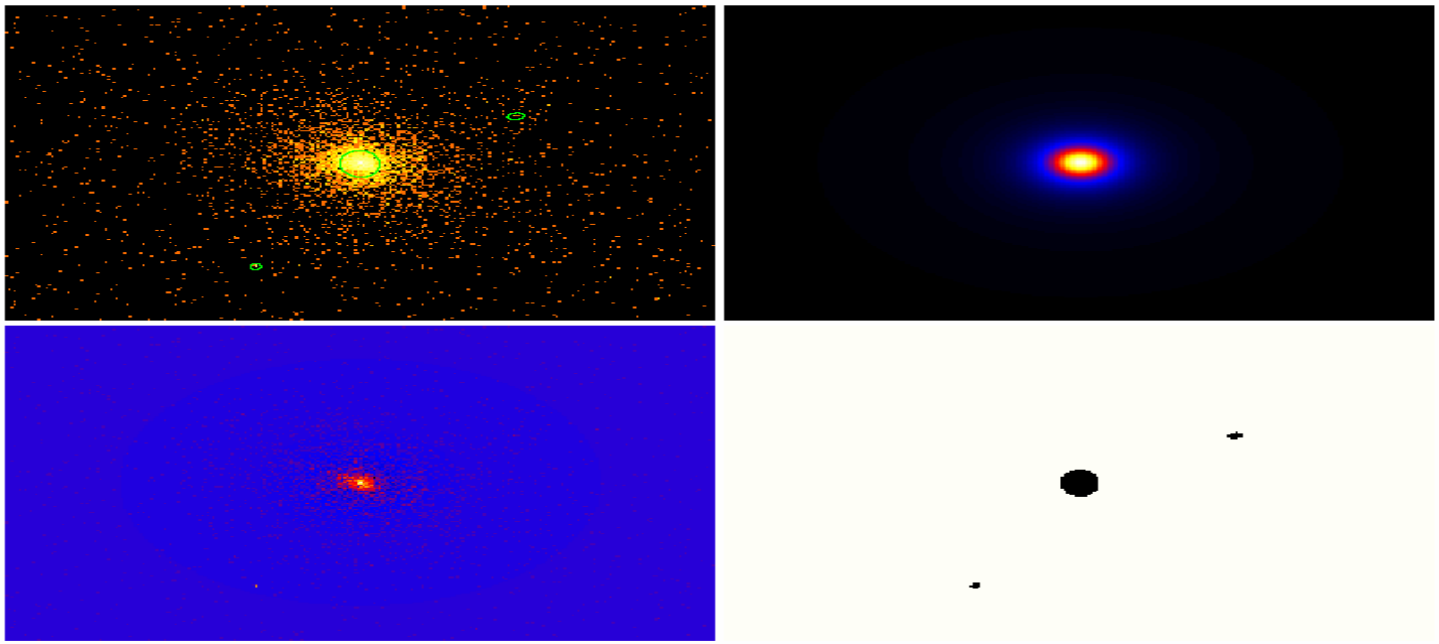


Figure 3: Example of a two-dimensional integer counts spectrum. *Top Left: Chandra* ACIS-S data of X-ray cluster MS 2137.3-2353, with *ds9* source regions superimposed. Top Right: Best-fit of a two-dimensional beta model to the filtered data. *Bottom Left:* Residuals (in units of σ) of the best fit. *Bottom Right:* The applied filter; the data within the ovals were excluded from the fit.

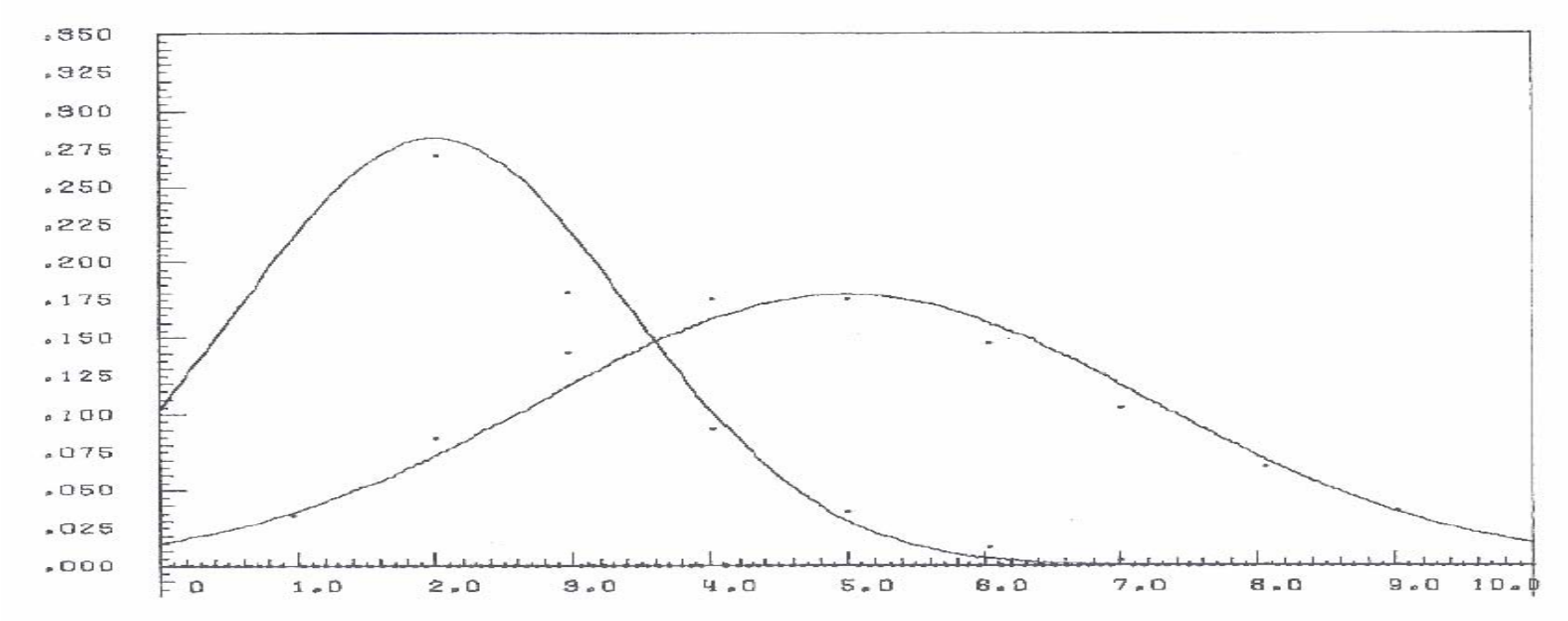


Figure 4: Comparison of Poisson distributions (dotted) of mean $\mu = 2$ and 5 with normal distributions of the same mean and variance (Eadie *et al.* 1971, p. 50).

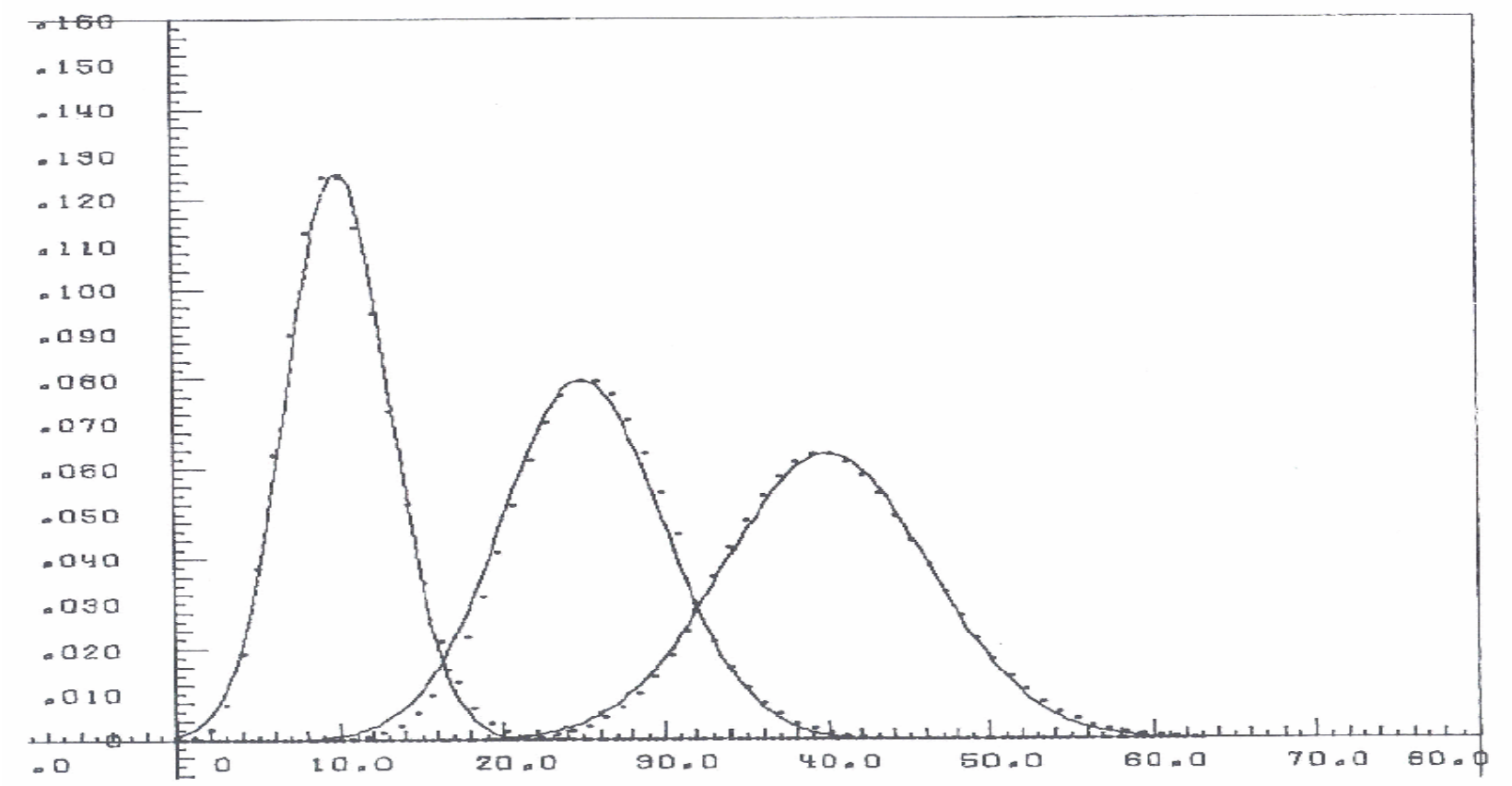


Figure 5: Comparison of Poisson distributions (dotted) of mean μ = 10, 25 and 40 with normal distributions of the same mean and variance (Eadie *et al.* 1971, p. 50).

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:

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

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 *Sherpa*); 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.

$$D_i! = \sqrt{2\pi D_i} D_i^{D_i} e^{-D_i}$$

$$\Rightarrow \mathcal{L}_i \approx \frac{1}{\sqrt{2\pi D_i}}\left(\frac{M_i}{D_i}\right)^{D_i} e^{D_i - M_i}$$

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

$$L_i = \log\mathcal{L}_i \approx -\frac{1}{2}\log(2\pi D_i) + D_i\log(\frac{M_i}{D_i}) + D_i - M_i \approx -\frac{1}{2}\log(2\pi D_i) + D\int \log(1+\frac{\epsilon}{\sqrt{D_i}}) - \epsilon\sqrt{D_i}$$

$$\approx -\frac{1}{2}\log(2\pi D_i) + D_i\left(\frac{\epsilon}{\sqrt{D_i}} - \frac{\epsilon^2}{2D_i} + \frac{\epsilon^3}{3D_i^{3/2}} - \cdots\right) - \epsilon\sqrt{D_i}$$

$$\approx -\frac{1}{2}\log(2\pi D_i) - \frac{\epsilon^2}{2} + \mathcal{O}(\frac{\epsilon^3}{\sqrt{D_i}})$$

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

[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

– $D_i >> 1$ 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 χ^2_{obs} will be sampled from the χ^2 distribution for N - P degrees of freedom:

$$p\left(\chi^2 \mid N - P\right) = \frac{1}{2\,\Gamma\left(\frac{N-P}{2}\right)}\left(\frac{\chi^2}{2}\right)^{\frac{N-P}{2}-1} e^{-\frac{\chi^2}{2}} .$$

Note that if N - P = 1, the χ^2 distribution diverges, while in the limit $N - P \to \infty$, it asymptotically approaches a Gaussian distribution with mean N - P and variance 2(N - P). Also note that if $P \geq N$, then the χ^2 distribution cannot be defined – you are doing something very wrong if you have more model parameters than data bins!

Validity of the χ^2 Statistic

So, when can you safely use the χ^2 statistic instead of the maximum likelihood in your fits?

This is a trick question - the answer is always. That's because you can run simulations to determine the distribution from which your observed value of χ^2 is sampled to quantitatively assess your fit. However, the whole reason one uses the χ^2 statistic is to avoid time-consuming simulations (and to use the χ^2 distribution directly)!

That said, the rules:

- A general rule-of-thumb says that χ^2 is sampled from the χ^2 distribution if there is a minimum of five counts in every bin. (But there is no standard reference for this in the literature, and the more counts, the better!)
- Also, the fit must be good!

Unfortunately, bad fits are common, even necessary, in X-ray astronomy; one example is the fit of a continuum model to data exhibiting an obvious (emission or absorption) line, done in an attempt to quantify how well the line is detected (a issue we'll return to later, when discussing model comparison). Inferences made using such fits *can be suspect!*

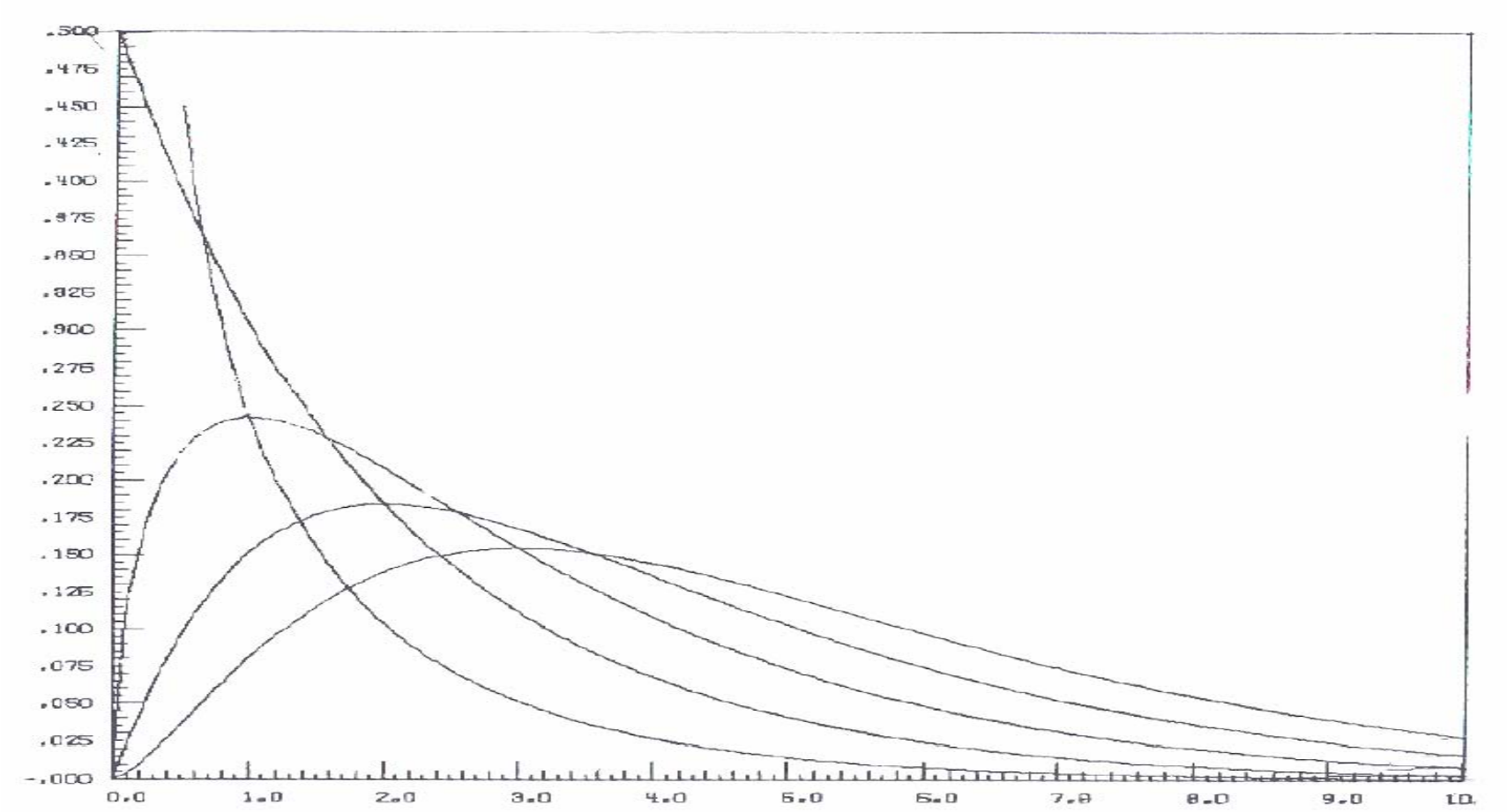


Figure 6: Examples of the χ^2 distribution for $v = N - P = 1, 2, 3, 4$, and 5 (Eadie *et al.* 1971, p. 64).

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:

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

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

$$\alpha_{\chi^2} = \int_{\chi^2_{obs}}^{\infty} d\chi^2 \, p(\chi^2 \mid N - P)$$

$$= \frac{1}{2\,\Gamma\left(\frac{N-P}{2}\right)} \int_{\chi^2_{obs}}^{\infty} d\chi^2 \left(\frac{\chi^2}{2}\right)^{\frac{N-P}{2}-1} e^{-\frac{\chi^2}{2}} .$$

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.,*

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

  - A "good" fit has $\chi^2_{obs,r} \approx 1$.
  - If $\chi^2_{obs,r} \to 0$ the fit is "too good" -- which means (1) the errorbars are too large, (2) χ^2_{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!

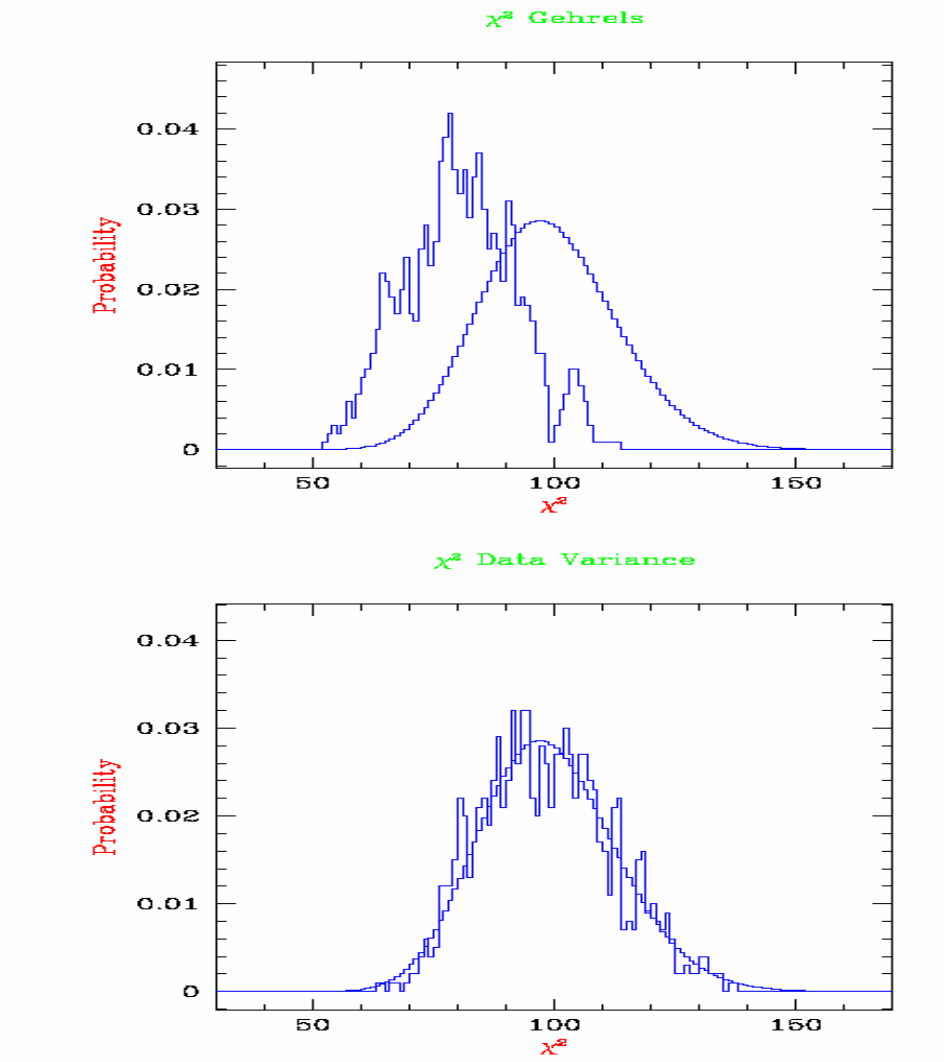


Figure 7: Comparison of the distributions of 500 sampled values of χ^2 versus the expected distribution for 99 degrees of freedom. Top: χ^2 with Gehrels variance. Bottom: χ^2 with data variance.

Statistical Issues: Background Subtraction

- A typical "dataset" may contain multiple spectra, one 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 $L_B \times L_{\mathrm{S+B}}$ or minimize $\chi^2_{\mathrm{B}} + \chi^2_{\mathrm{S+B}}$.
- 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.

```
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
```

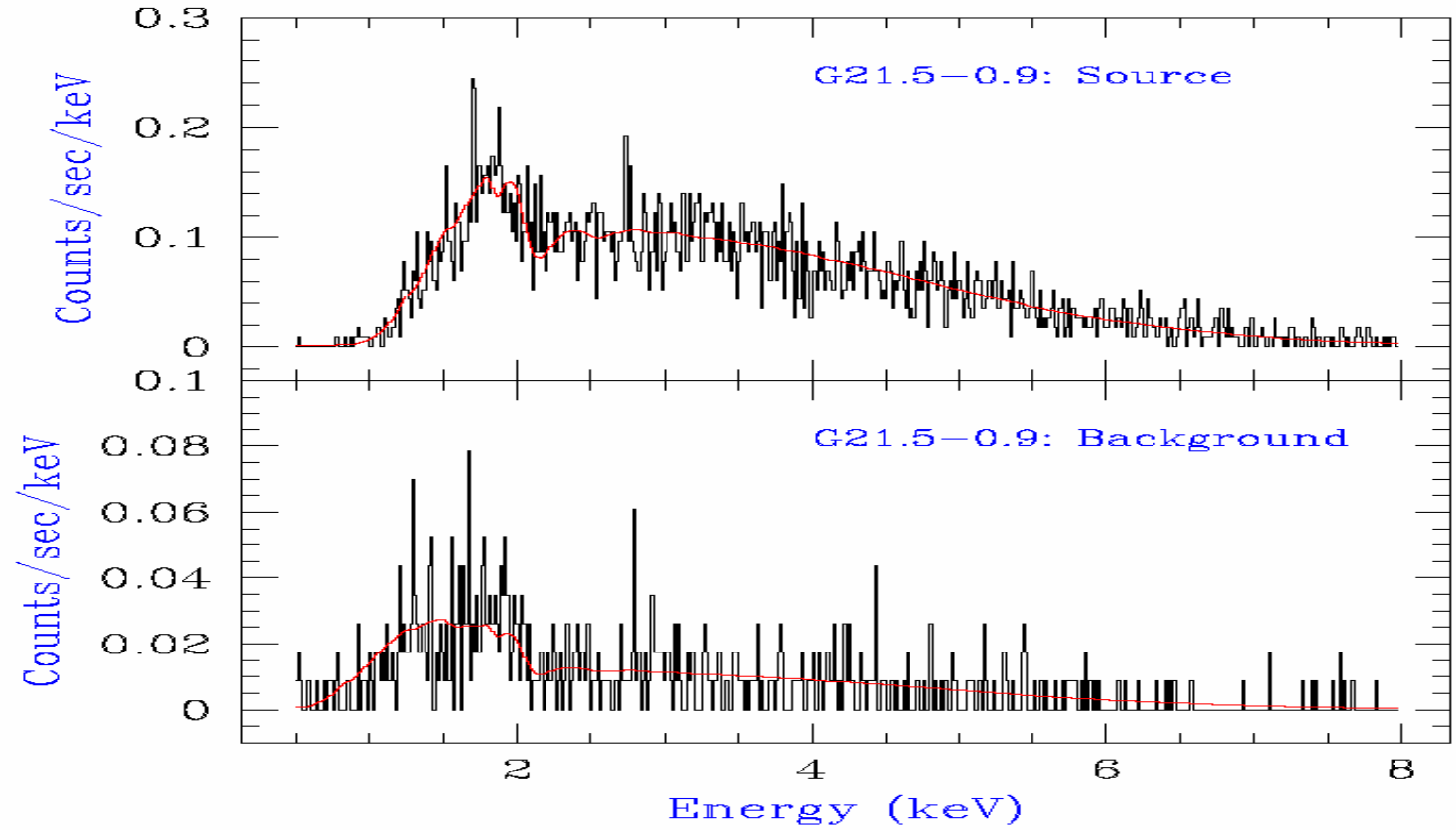


Figure 8: *Top:* Best-fit of a power-law times galactic absorption model to the source spectrum of supernova remnant G21.5-0.9. *Bottom:* Best-fit of a separate power-law times galactic absorption model to the background spectrum extracted for the same source.

Statistical Issues: Background Subtraction

- 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*:

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

  - 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: 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 <= 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

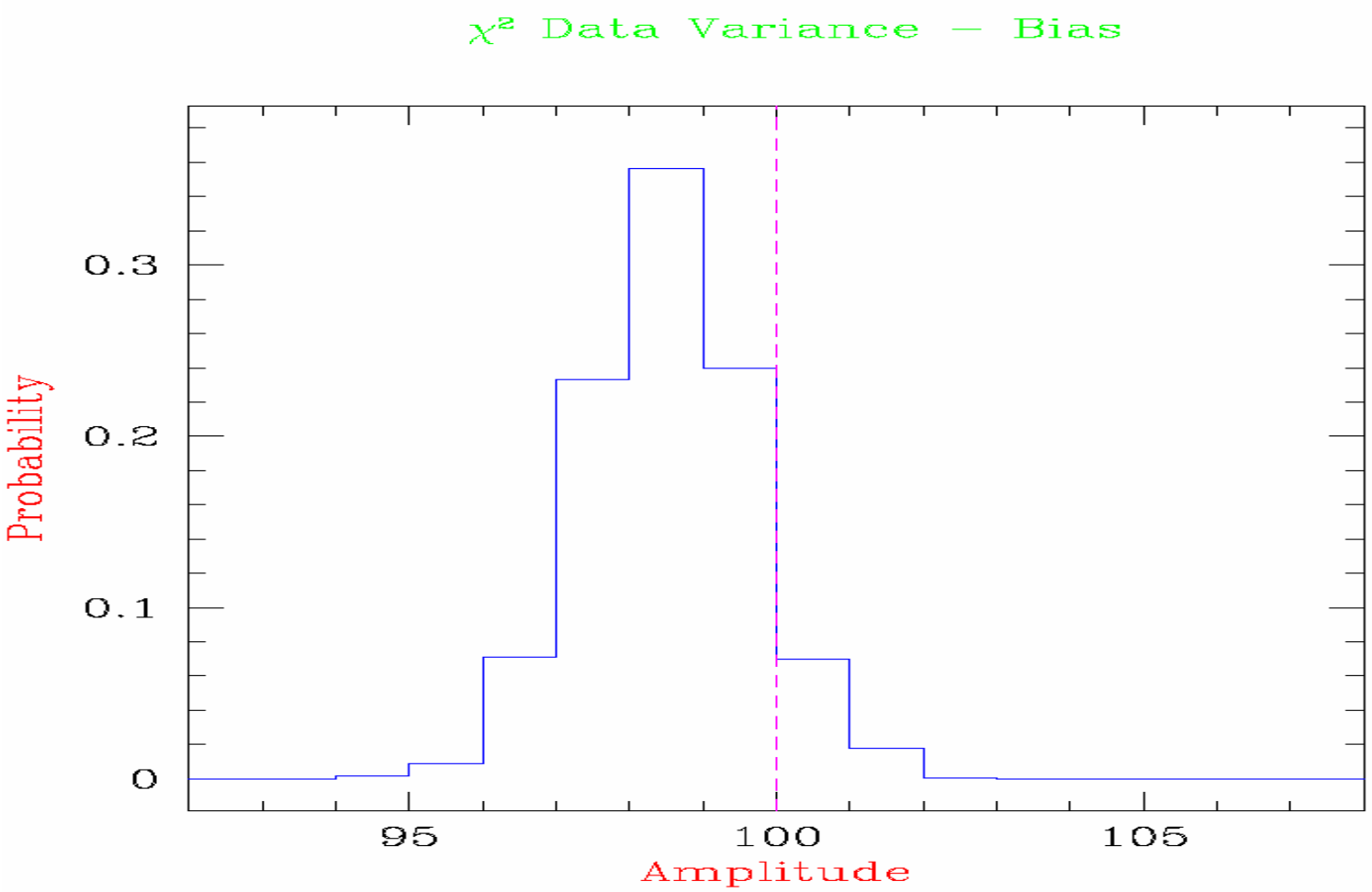


Figure 9: A demonstration of bias. Five hundred datasets are sampled from a constant model with amplitude 100 and then are fit with the same constant amplitude model, using χ^2 with data variance. The mean of the distribution of fit amplitude values is not 100, as it would be if the statistic were an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

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 $M\left(\hat{\theta}\right)$ 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 $M\left(\hat{\theta}\right)$ 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

$$p(\theta \mid D) = p(\theta)\frac{p(D \mid \theta)}{p(D)}$$

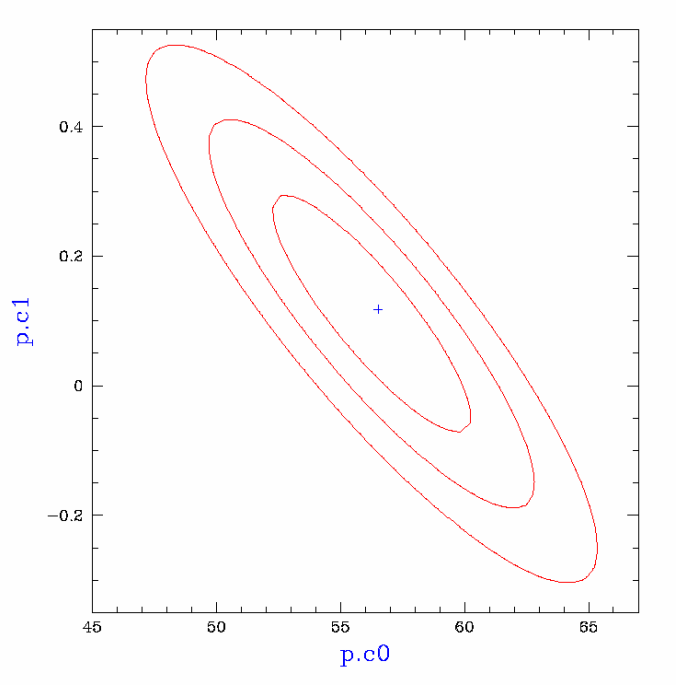
where

- $p(D|\theta)$ 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!)

Statistics II:

Model Comparison and Parameter Estimation



Peter Freeman

Harvard-Smithsonian Center for Astrophysics

Now, Shifting Gears…

A model *M* has been fit to dataset *D* and either the maximum of the likelihood function $\mathcal{L}_{max}$, the minimum of the χ^2 statistic χ^2_{min}, or the mode of the posterior distribution $p(\hat{\theta} \mid D)$ has been determined. What comes next?

- *Model Comparison.* The determination of which of a suite of models (*e.g.,* blackbody, power-law, *etc.*) best represents the data.
- *Parameter Estimation.* The characterization of the sampling distribution for each best-fit model parameter (*e.g.,* blackbody temperature and normalization), which allows the errors (*i.e.,* standard deviations) of each parameter to be determined.
- *Publication!*

Here, we cannot ignore the frequentist/Bayesian divide. Hence we will discuss how frequentists and Bayesians would complete these tasks, separately…

Frequentist Model Comparison

Two models, M_0 and M_1, have been fit to D. M_0, the "simpler" of the two models (generally speaking, the model with fewer free parameters) is the *null hypothesis*.

A frequentist would compare these models by:

- constructing a test statistic T from the best-fit statistics of each fit (*e.g.*, $\Delta\chi^2 = \chi^2_0 - \chi^2_1$);
- determining each sampling distributions for T, $p(T \mid M_0)$ and $p(T \mid M_1)$;
- determining the *significance*, or Type I error, the probability of selecting M_1 when M_0 is correct:

$$\alpha = \int_{T_{\rm obs}}^{\infty} dT\, p(T \mid M_0);$$

- and determing the *power*, or Type II error, which is related to the probability β of selecting M_0 when M_1 is correct:

$$1 - \beta = \int_{T_{\rm obs}}^{\infty} dT\, p(T \mid M_1).$$

$\Rightarrow$If α is smaller than a pre-defined threshold (≤ 0.05, or $\leq 10^{-4}$, *etc.*, with smaller thresholds used for more controversial alternative models), then the frequentist rejects the null hypothesis.

$\Rightarrow$If there are several model comparison tests to choose from, the frequentist uses the most powerful one!

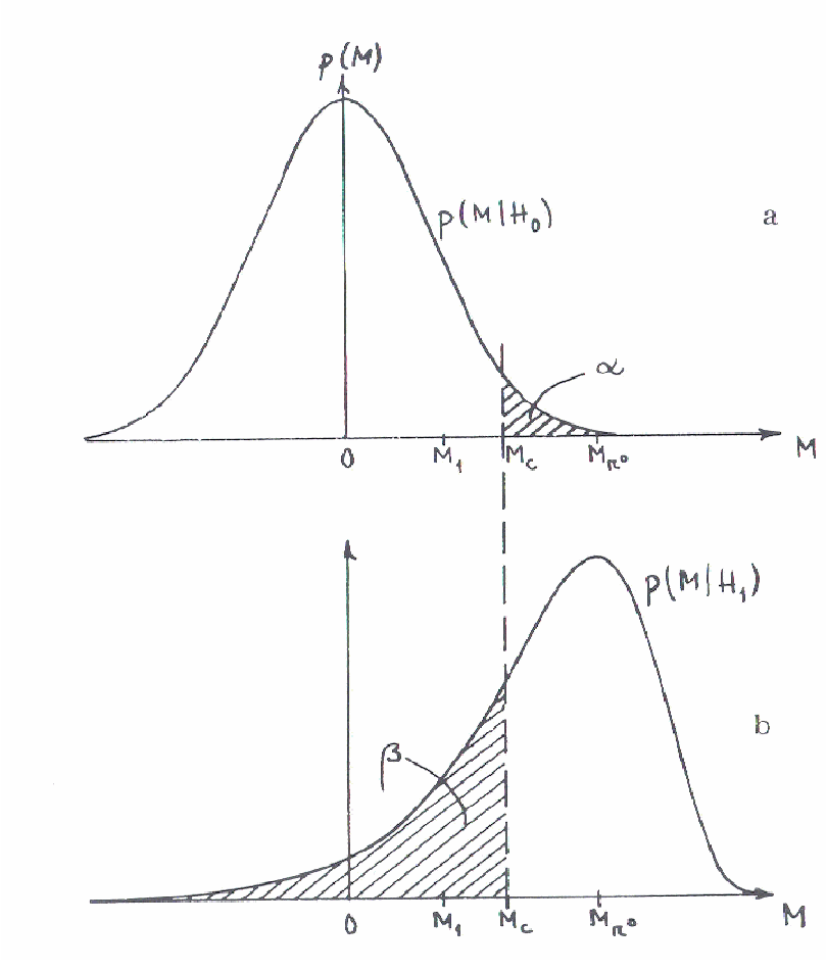


Figure 1: Comparison of distributions $p(T \mid M_0)$ (from which one determines the significance α) and $p(T \mid M_1)$ (from which one determines the power of the model comparison test $1 - \beta$) (Eadie *et al.* 1971, p.217)

Frequentist Model Comparison

Standard frequentist model comparison tests include:

- The χ^2 Goodness-of-Fit (GoF) test:

$$\alpha_{\chi^2} = \int_{\chi^2_{\min,0}}^{\infty} d\chi^2 \, p(\chi^2 \mid N - P_0) = \frac{1}{2\Gamma(\frac{N-P_0}{2})} \int_{\chi^2_{\min,0}}^{\infty} d\chi^2 (\frac{\chi^2}{2})^{\frac{N-P_0}{2}-1} e^{-\frac{\chi^2}{2}} .$$

- The Maximum Likelihood Ratio (MLR) test:

$$\alpha_{\chi^2 MLR} = \int_{\Delta\chi^2}^{\infty} d\chi^2 \, p(\Delta\chi^2 \mid \Delta P),$$

where ΔP is the number of additional freely varying model parameters in model M_1.

- The F-test:

$$\alpha_F = \int_F^{\infty} dF \; p(F \mid \Delta P, N - P_1 = I_{\frac{N-P_1}{N-P_1+(\Delta P)F}} (\frac{N-P_1}{2}, \frac{\Delta P}{2}),$$

where P_1 is the total number of thawed parameters in model M_1, I is the incomplete beta function, and F is the F-statistic

$$F = \frac{\Delta\chi^2}{\Delta P} / \frac{\chi_1^2}{(N - P_1)} .$$

These are standard tests because they allow estimation of the significance without time-consuming simulations!

Frequentist Model Comparison

Notes and caveats regarding these standard tests:

- The GoF test is an "alternative-free" test, as it does not take into account the alternative model M_1. It is consequently a *weak* (*i.e.,* not powerful) model comparison test and should not be used!
- Only the version of *F*-test which generally has the greatest power is shown above: in principle, one can construct three *F* statistics out of χ_0^2, χ_1^2, and $\Delta\chi^2$
- The MLR ratio test is generally the most powerful for detecting emission and absorption lines in spectra.

But the most important caveat of all is that…

Frequentist Model Comparison

The *F* and MLR tests are commonly misused by astronomers! There are two important conditions that must be met so that an estimated derived value α is actually correct, *i.e.,* so that it is an accurate approximation of the tail integral of the sampling distribution (Protassov *et al.* 2001):

- M_0 must be *nested* within M_1, *i.e.,* one can obtain M_0 by setting the extra ΔP parameters of M_1 to default values, often zero; and
- those default values may not be on a parameter space boundary.

The second condition may not be met, *e.g.,* when one is attempting to detect an emission line, whose default amplitude is zero and whose minimum amplitude is zero. Protassov *et al.* recommend Bayesian posterior predictive probability values as an alternative, but a discussion of this topic is beyond the scope of this class.

If the conditions for using these tests are not met, then they can still be used, but the significance must be computed via Monte Carlo simulations.

Bayesian Model Comparison

In the previous class, we showed how Bayes' theorem is applied in model fits. It can also be applied to model comparison:

$$p(M \mid D) = p(M)\frac{p(D \mid M)}{p(D)}.$$

- $p(M)$ is the prior probability for M;
- $p(D)$ is an ignorable normalization constant; and
- $p(D \mid M)$ is the average, or global, likelihood:

$$p(D \mid M) = \int d\theta\, p(\theta \mid M)\, p(D \mid M, \theta)$$
$$= \int d\theta\, p(\theta \mid M)\, \mathcal{L}(M, \theta).$$

In other words, it is the (normalized) integral of the posterior distribution over all parameter space. Note that this integral may be computed numerically, by brute force, or if the likelihood surface is approximately a multi-dimensional Gaussian (*i.e.* if $\mathcal{L}$ α exp[-$\chi^2/2$]), by the *Laplace approximation:*

$$p(D \mid M) = p(\hat{\theta} \mid M)(2\pi)^{P/2}\sqrt{\det C}\;\mathcal{L}_{\max},$$

where C is the covariance matrix (estimated numerically at the mode).

Bayesian Model Comparison

To compare two models, a Bayesian computes the odds, or odd ratio:

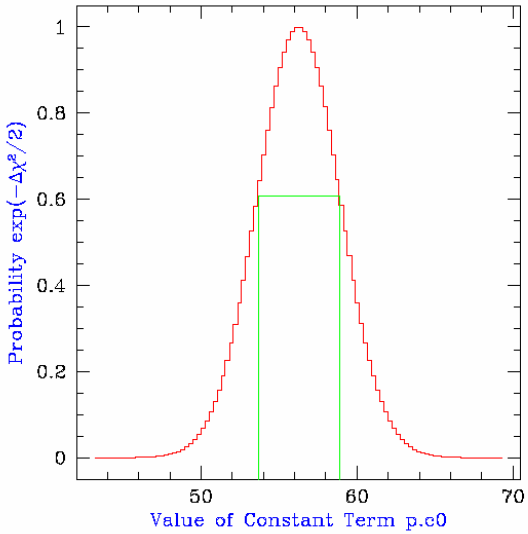
$$O_{10} = \frac{p(M_1 \mid D)}{p(M_0 \mid D)} = \frac{p(M_1)\,p(D \mid M_1)}{p(M_0)\,p(D \mid M_0)} = \frac{p(M_1)}{p(M_0)} B_{10} \,,$$

where B_{10} is the *Bayes factor*. When there is no *a priori* preference for either model, $B_{10} = 1$ of one indicates that each model is equally likely to be correct, while $B_{10} \geq 10$ may be considered sufficient to accept the alternative model (although that number should be greater if the alternative model is controversial).

Parameter Estimation

One should speak of *confidence* or *credible intervals* or *regions* rather than "errors."

- A frequentist derives confidence intervals and regions.
- A Bayesian derives credible intervals and regions.
- An interval is a range (or ranges) of values of a parameter θ that has probability $p_{\rm int}$ of containing the parameter's true value θ_o. (A region is simply the multi-dimensional analogue of an interval.)
- An infinite number of intervals can be defined for a given parameter: here, we'll speak of intervals that contain the *most probable* parameter values.



Parameter Estimation

Instead of the integrated probability $p_{\rm int}$, many speak of "numbers of σ." One can convert from $n\sigma$ to $p_{\rm int}$ using the following equation:

$$p_{\rm int} = \frac{1}{\sqrt{2\pi}\sigma}\int_{-n\sigma}^{+n\sigma} dx \exp(-\frac{x^2}{2\sigma^2}) = \mathrm{erf}(\frac{n}{\sqrt{2}})$$

$p_{\rm int}$	σ
68.3%	1.0
90.0%	1.6
95.5%	2.0
99.0%	2.6
99.7%	3.0

Note: this conversion between $p_{\rm int}$ and σ, while strictly true only if the sampling distribution is a one-dimensional Gaussian, is used by many astronomers in casual conversation regardless of the actual distribution shape or dimensionality.

Parameter Estimation

- Tables showing $\Delta\chi^2$ as a function of integrated probability $p_{\rm int}$ and number of degrees of freedom $\nu = N - P$ can cause confusion. For instance:
  - "I have two free parameters in my model. Hence I should compute 68.3% confidence intervals for each parameter using $\Delta\chi^2 = 2.30$, right?"
  - "No."

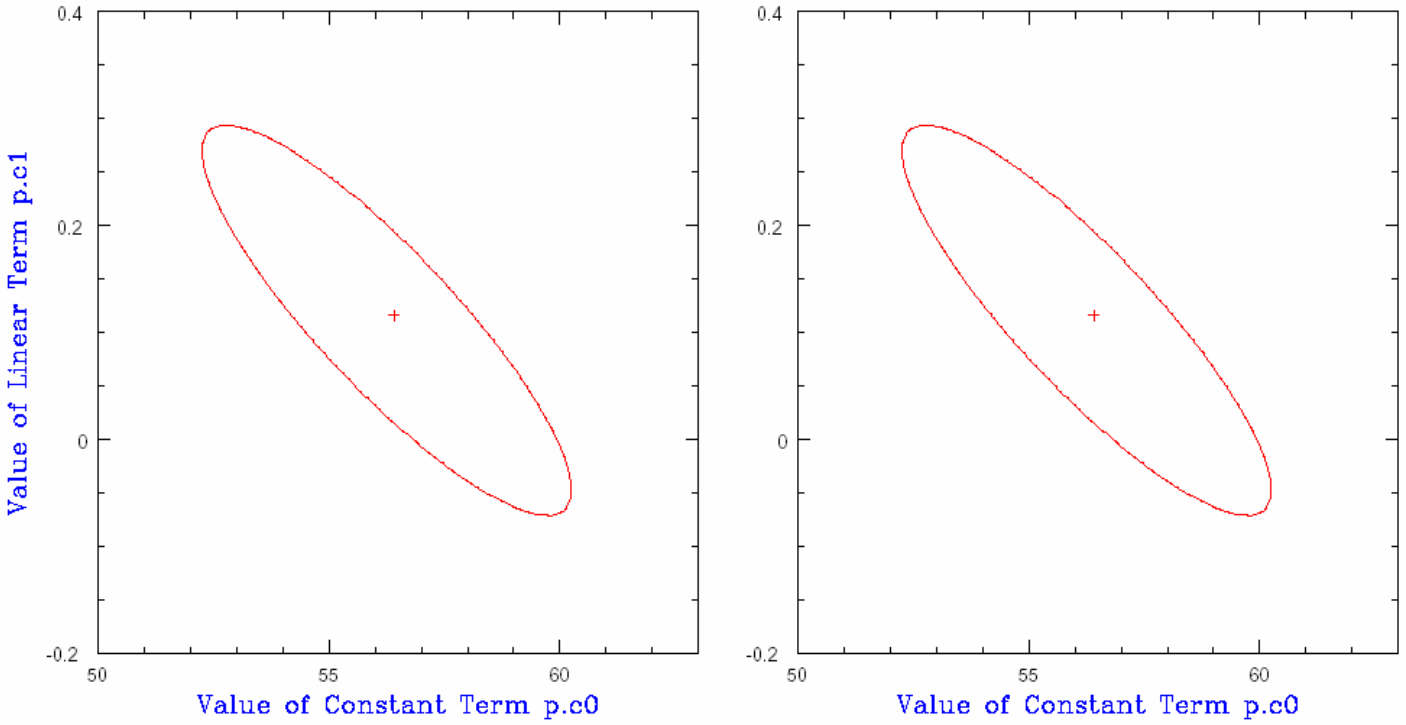
	ν					
$p_{\rm int}$	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom (Based on Press et al. 1986, p. 536.)

- To find the $n\sigma$ confidence interval for one parameter, use $\Delta\chi^2$ for $\nu = 1$ (or n^2).
- To find the *$n\sigma$ joint confidence region* for m parameters, use $\Delta\chi^2$ for $\nu = m$.
- To find either an interval or region using the likelihood function $\mathcal{L}$, use $\Delta\log\mathcal{L} = \Delta\chi^2/2$.

Parameter Estimation

Never project a (properly estimated) region onto a parameter axis to estimate an interval! This always *over-estimates* the size of the interval.



Frequentist Parameter Estimation

To determine confidence intervals and regions, a frequentist generally must simulate and fit new datasets to determine the sampling distributions for each model parameter.

- If the true parameter values are unknown (which is usually the case), then a grid of model parameter values should be constructed, with a large number of datasets sampled at each grid point.
- But the usual choice is to appeal to asymptotic behavior and sample datasets using $M(\hat{\theta})$. This method may only be useful in limited circumstances, as >= 100 datasets should be sampled and fit for accurate results.

Frequentist Parameter Estimation

One can estimate confidence intervals without having to use simulations if the χ^2 or log $\mathcal{L}$ surface in parameter space is "well-behaved," *i.e.,* if

- the surface is approximately shaped like a multi-dimensional paraboloid; and
- the best-fit point is sufficiently far from parameter-space boundaries.

Three common ways of determining $n\sigma$ intervals are:

- Varying a parameter's value, *while holding the values of all other parameters at their best-fit values,* until

$$\chi^2 = \chi_o^2 + n^2; \text{ or } \log\mathcal{L} = \log\mathcal{L}_o - \frac{n^2}{2};$$

- the same as above, but *allowing the values of all other parameters are allowed to float to new best-fit values;* and
- computing $n\sqrt{C_{i,i}}$, where the covariance matrix $C_{i,j} = I^{-1}_{i,j}$ and I, the information matrix computed at the best-fit point, is

$$I_{i,j} \equiv \frac{1}{2}\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j} \text{ or } \frac{\partial^2 \log\mathcal{L}}{\partial\theta_i\partial\theta_j}.$$

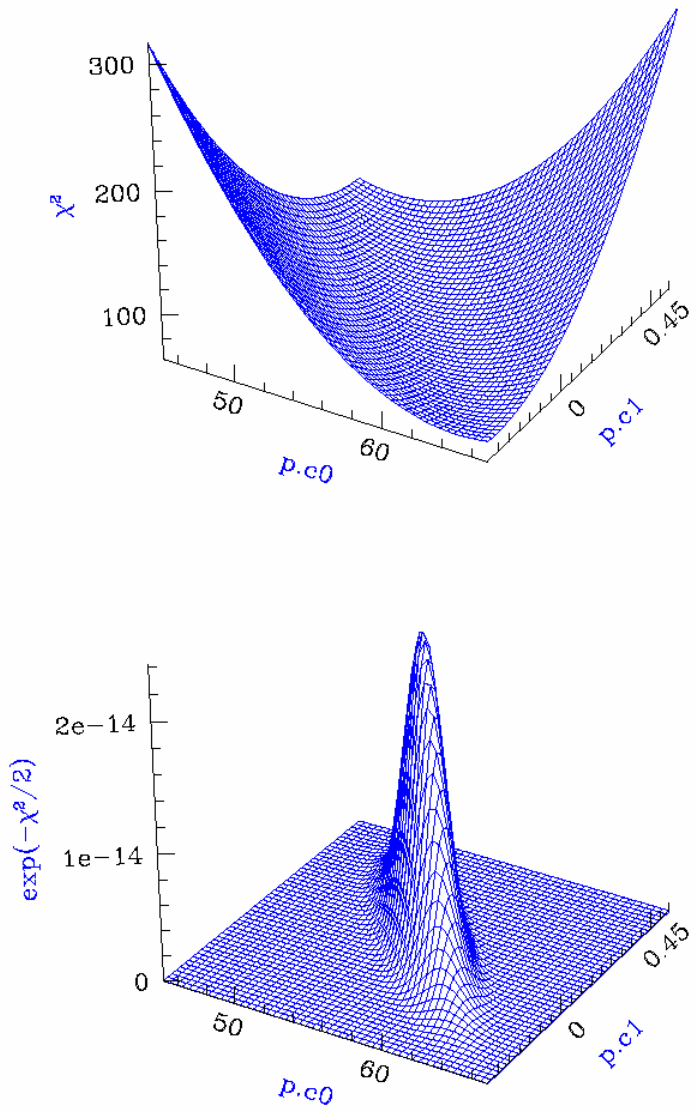


Figure 2: Example of a "well-behaved" statistical surface in parameter space, viewed as a multi-dimensional paraboloid (χ^2, *top*), and as a multi-dimensional Gaussian ($\exp(-\chi^2/2) \approx \mathcal{L}$, *bottom*).

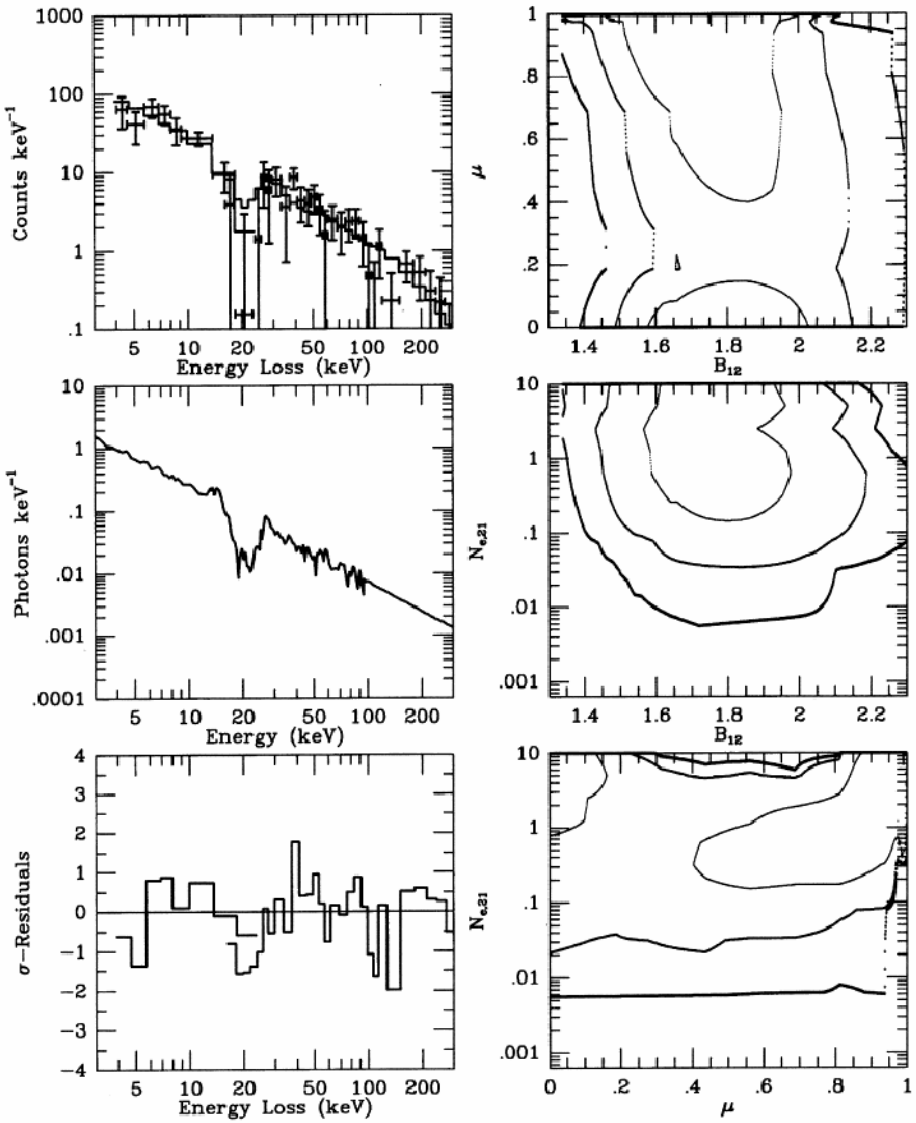


Figure 3: On the right, 1, 2, and 3σ contours determined for a statistical surface that is not "well-behaved" in parameter space. With such a surface, rigorous parameter estimation involves simulations (frequentist approach) or numerical integration of the surface (Bayesian approach). From Freeman *et al.* (1999).

Frequentist Parameter Estimation

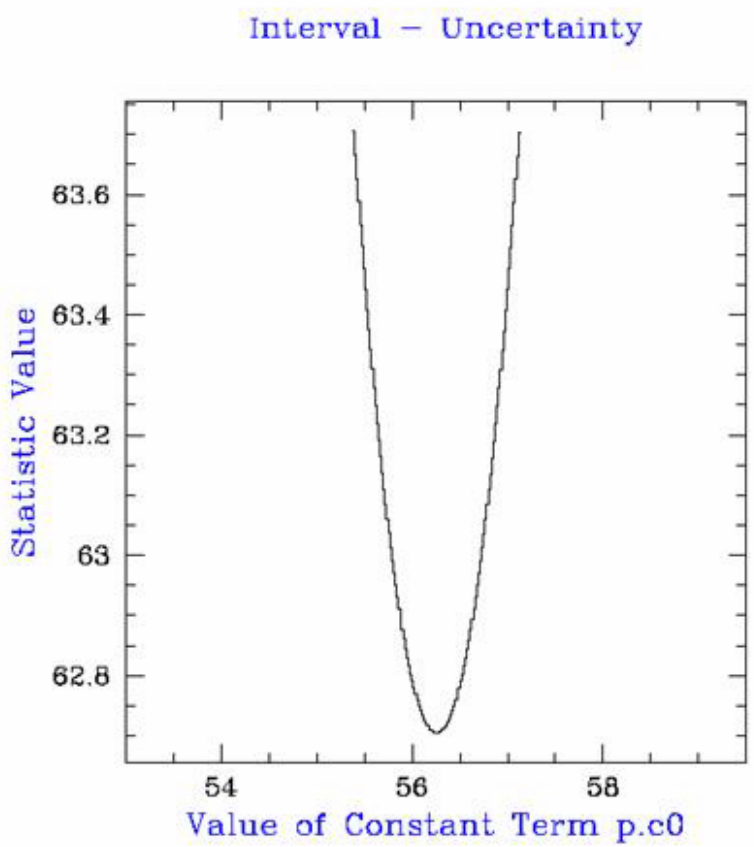
Things to keep in mind about these confidence interval estimators (dubbed `UNCERTAINTY`, `PROJECTION`, and `COVARIANCE` in *Sherpa*, respectively):

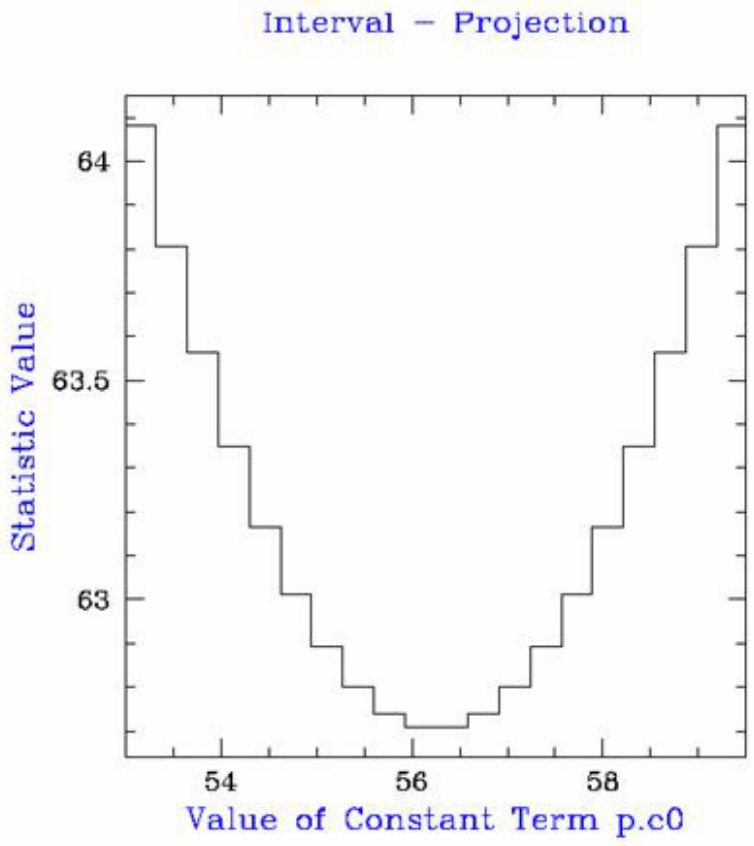
- The first method will *always* underestimate the interval if the value of the parameter of interest is correlated with other model parameter values.
- The second method (which is relatively slow) is in a rigorous sense no more accurate than the third method (which is fast), but it does provide a means of visualizing the statistical surface.
- A statistical surface is "well-behaved" if the second and third methods give the same interval estimates.
- The condition that the best-fit point be sufficiently far from parameter-space boundaries means that these methods are *not* appropriate for determining upper or lower limits.

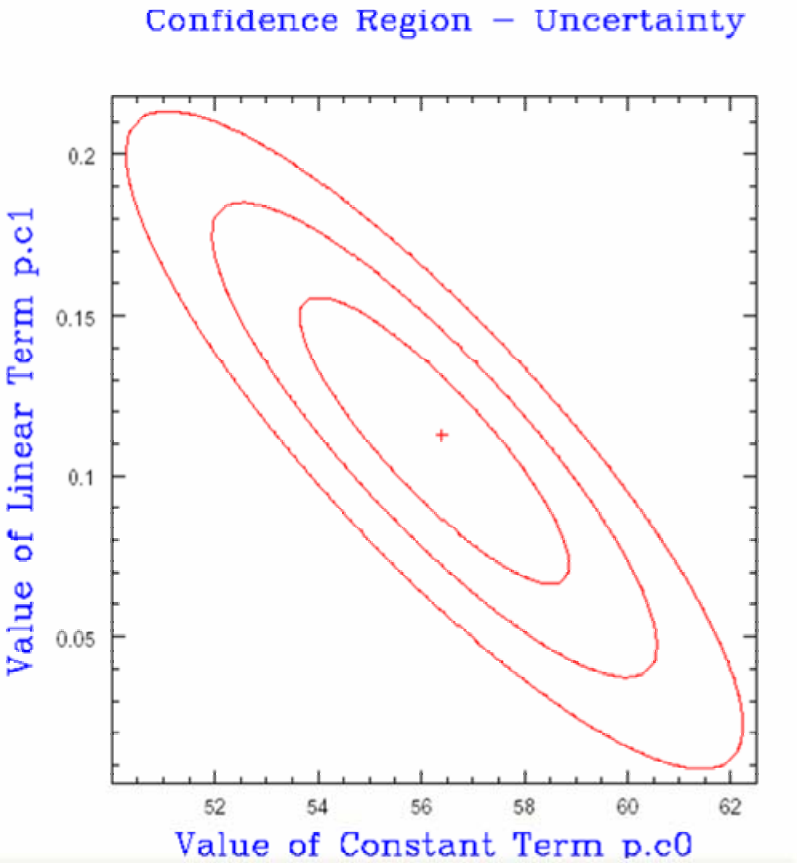
Example with a Well-Behaved Parameter Space

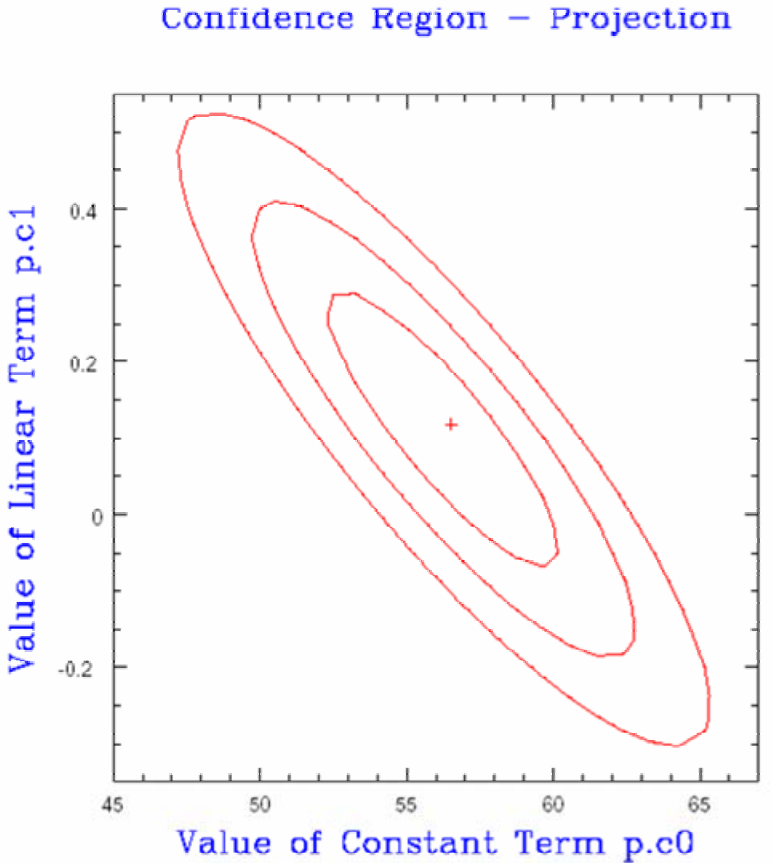
```
sherpa> fit
 powll: v1.2
 powll:   initial function value =   8.22297E+01
 powll:     converged to minimum =   6.27050E+01 at iteration =         7
 powll:   final function value   =   6.27050E+01
           p.c0   56.2579
           p.c1   0.11117
           p.c2   -0.00119999

sherpa>  uncertainty
Computed for uncertainty.sigma = 1
      ------------------------------------------------------------
      Parameter Name     Best-Fit   Lower Bound     Upper Bound
      ------------------------------------------------------------
          p.c0           56.2579   -0.865564         +0.864461
          p.c1           0.11117   -0.0148228        +0.0148038
          p.c2       -0.00119999   -0.000189496      +0.000189222
```

```
sherpa> projection
Computed for projection.sigma = 1
      ---------------------------------------------------------
      Parameter Name      Best-Fit   Lower Bound    Upper Bound
      ---------------------------------------------------------
          p.c0             56.2579   -2.64465       +2.64497
          p.c1             0.11117   -0.120684      +0.120703
          p.c2         -0.00119999   -0.00115029    +0.00114976

sherpa> covariance
Computed for covariance.sigma = 1
      ---------------------------------------------------------
      Parameter Name     Best-Fit    Lower Bound      Upper Bound
      ---------------------------------------------------------
          p.c0             56.2579   -2.64786       +2.64786
          p.c1             0.11117   -0.121023      +0.121023
          p.c2         -0.00119999   -0.00115675    +0.00115675
```


Interval – Uncertainty
Statistic Value
63.6
63.4
63.2
63
62.8
54
56
58
Value of Constant Term p.c0


Interval – Projection
Statistic Value
64
63.5
63
54
56
58
Value of Constant Term p.c0


Confidence Region – Uncertainty
Value of Linear Term p.c1
0.2
0.15
0.1
0.05
52
54
56
58
60
62
Value of Constant Term p.c0


Confidence Region – Projection
Value of Linear Term p.c1
0.4
0.2
0
-0.2
45
50
55
60
65
Value of Constant Term p.c0

Bayesian Parameter Estimation

A Bayesian estimates credible intervals and regions by *marginalizing* (integrating) the parameter posterior distribution over the space of *nuisance* (uninteresting) parameters. For instance:

$$p(\theta_1 \mid D) = \int_{\theta_2} d\theta_2 \ldots \int_{\theta_P} d\theta_P \; p(\theta \mid D).$$

The central 68% of the distribution $p(\theta_1 \mid D)$ is the 1σ credible interval.

Marginalization may be done by brute-force integration or, for higher dimensional problems ($N \gtrsim 10$), by adaptive integration. However, if the statistical surface is "well-behaved," one can also estimate credible intervals using the Laplace Approximation:

$$p(\theta_1 \mid D) = p(\hat{\theta}_2, \ldots, \hat{\theta}_P)(2\pi)^{(P-1)/2} \times \sqrt{\det C(\theta_2, \hat{\theta}_2, \ldots, \hat{\theta}_P)}\, \mathcal{L}(\theta_1, \hat{\theta}_2, \ldots, \hat{\theta}_P).$$

If the values of parameter θ_1 is correlated with other parameter values, then when computing $p(\theta_1 \mid D)$, the values of parameters ($\theta_1, \cdots, \theta_P$) should be allowed to *float* to new best-fit values.

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:

- Loredo, 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\item Churazov, E., et al. 1996, ApJ 471, 673
- Gehrels, N. 1986, ApJ 303, 336\item Kearns, K., Primini, 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