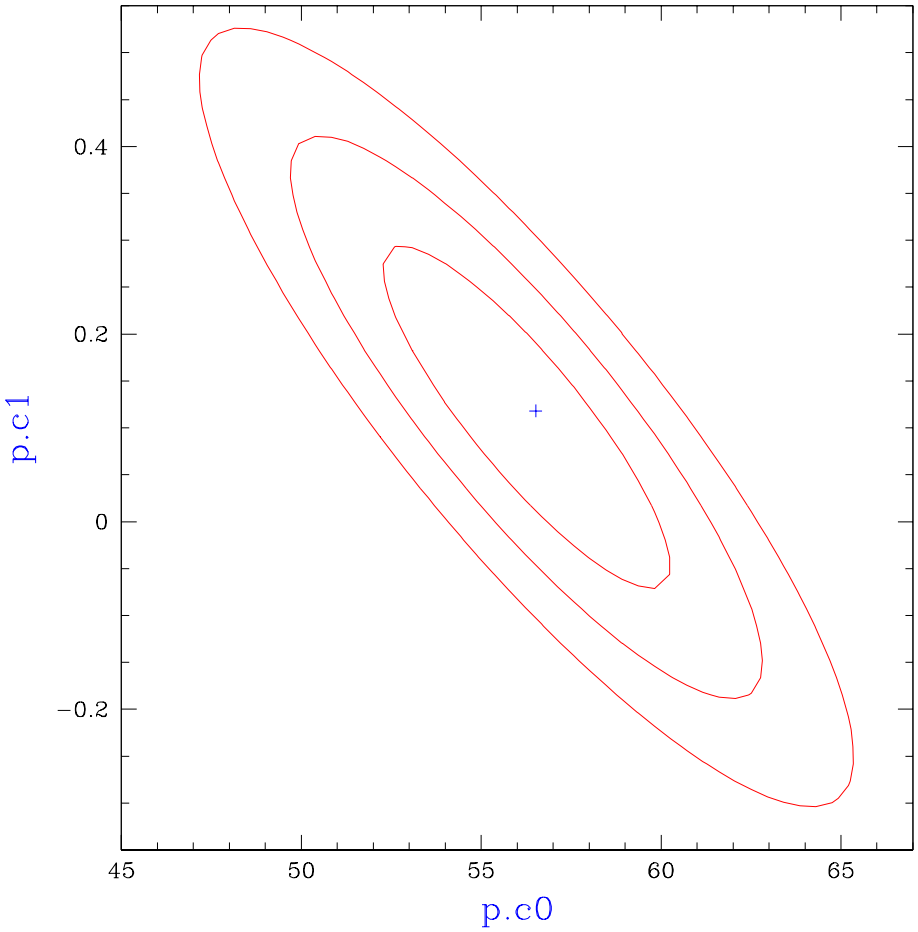
Statistics II:

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 $\mathcal{L}_{\rm max}$, the minimum of the χ^2 statistic $\chi^2_{\rm min}$, or the mode of the posterior distribution $p(\hat{\theta}|D)$ 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, $p(T|M_0)$ and $p(T|M_1)$;
- determining the *significance*, or Type I error, the probability of selecting M_1 when M_0 is correct:

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

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

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

$\Rightarrow$ 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!

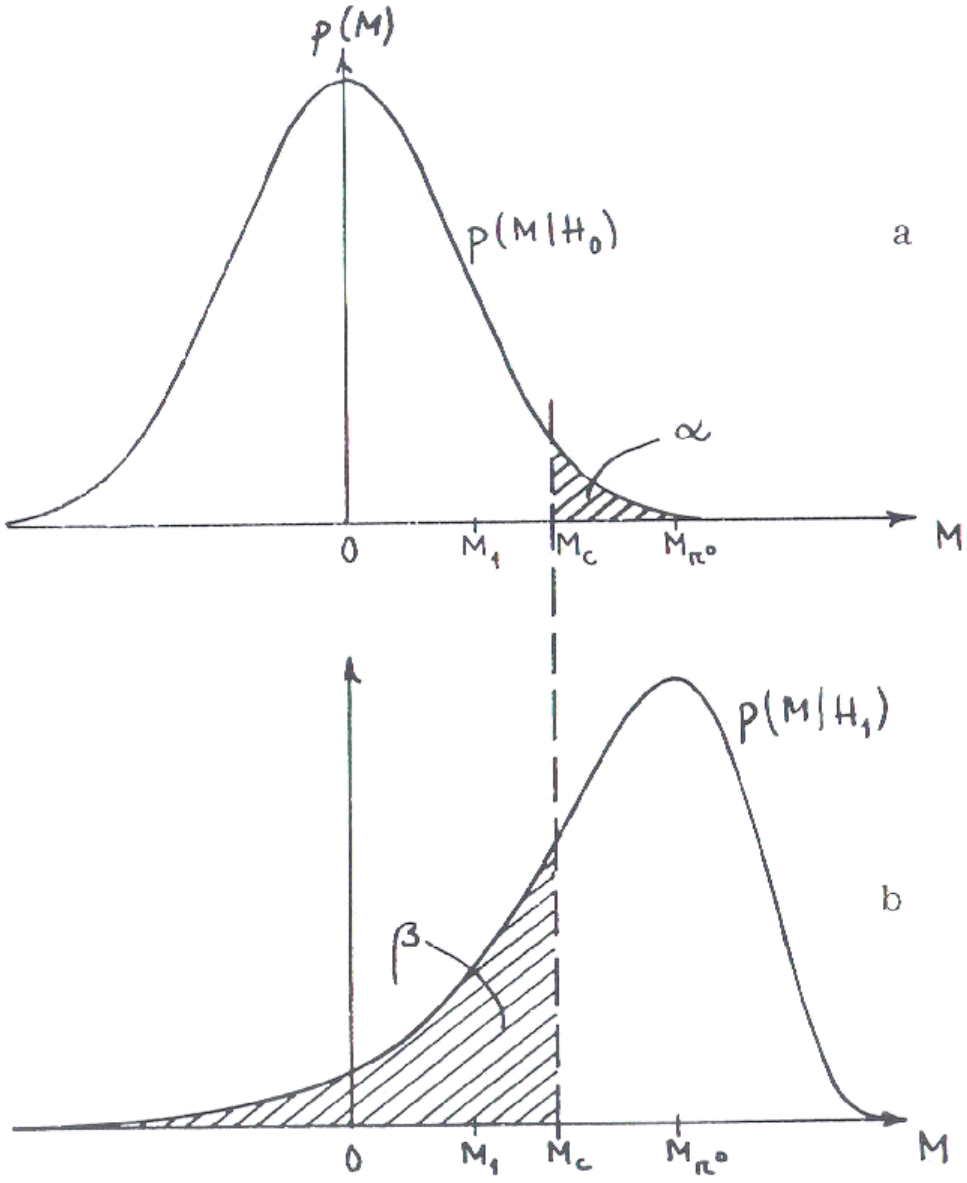


Figure 1: Comparison of distributions $p(T|M_0)$ (from which one determines the significance α) and $p(T|M_1)$ (from which one determines the power of the model comparison test $1-\beta$ (Eadie *et al.* 1971, p. 217).

Frequentist Model Comparison

Standard frequentist model comparison tests include:

- The χ^2 Goodness-of-Fit (GoF) test:

$$\begin{aligned} \alpha_{\chi^2} &= \int_{\chi^2_{\rm min,0}}^{\infty} d\chi^2 p(\chi^2|N-P_0) \\ &= \frac{1}{2\Gamma(\frac{N-P_0}{2})} \int_{\chi^2_{\rm min,0}}^{\infty} d\chi^2 \left(\frac{\chi^2}{2}\right)^{\frac{N-P_0}{2}-1} e^{-\frac{\chi^2}{2}} . \end{aligned}$$

- The Maximum Likelihood Ratio (MLR) test:

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

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

- The F-test:

$$\begin{aligned} \alpha_F &= \int_F^{\infty} dF\ p(F|\Delta P, N-P_1) \\ &= I_{\frac{N-P_1}{N-P_1+(\Delta P)F}} \left(\frac{N-P_1}{2}, \frac{\Delta P}{2}\right) , \end{aligned}$$

  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

$$F = \frac{\Delta\chi^2}{\Delta P} \Big/ \frac{\chi^2_1}{(N-P_1)} .$$

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 the F-test which generally has the greatest power is shown above: in principle, one can construct three F statistics out of χ^2_0, χ^2_1, 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:

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

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

$$\begin{aligned} p(D|M) &= \int d\theta p(\theta|M)p(D|M,\theta) \\ &= \int d\theta p(\theta|M)\mathcal{L}(M,\theta)\,. \end{aligned}$$

  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 $\mathcal{L} \propto \exp[-\chi^2/2]$), by the *Laplace approximation*:

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

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

Bayesian Model Comparison

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

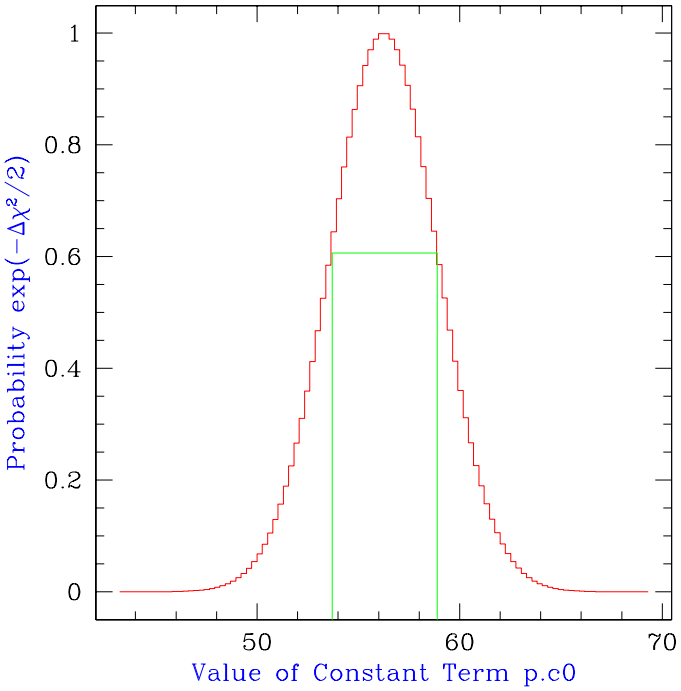
$$\begin{aligned} O_{10} &= \frac{p(M_1|D)}{p(M_0|D)} \\ &= \frac{p(M_1)p(D|M_1)}{p(M_0)p(D|M_0)} \\ &= \frac{p(M_1)}{p(M_0)} B_{10}\,, \end{aligned}$$

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 Bayesians 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.)
- A 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:

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

$p_{\rm int}$	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."

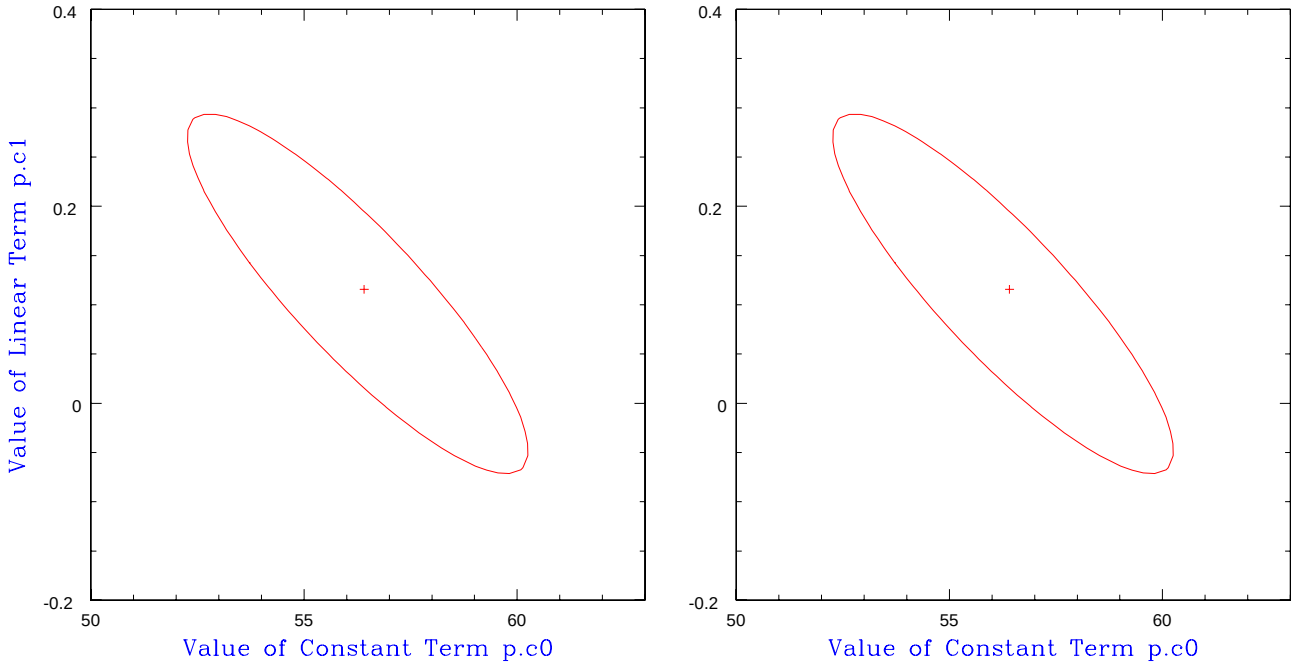
	ν					
$p_{\rm int}$	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

$\Delta\chi^2$ 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 *overestimates* 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 $\gtrsim 100$ 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
  - $\chi^2 = \chi^2_o + n^2$; 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

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

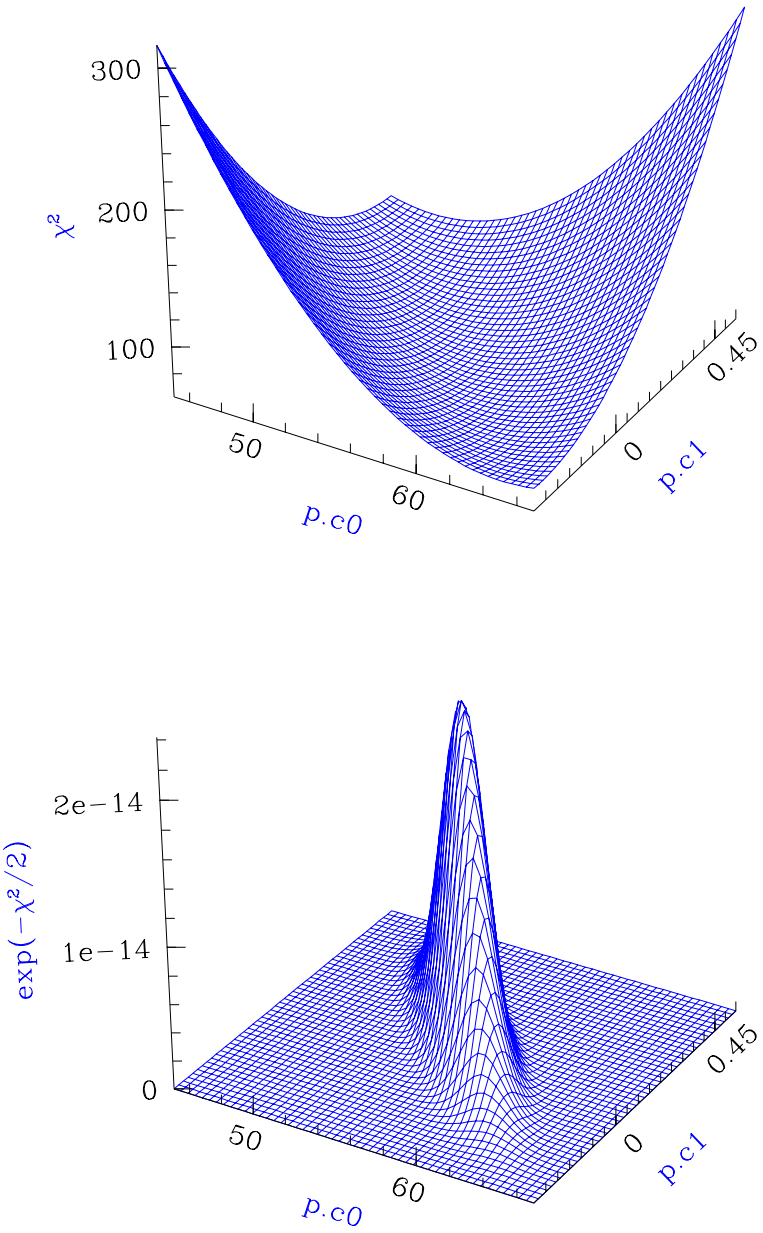


Figure 2: Example of a "well-behaved" statistical surface in parameter space, viewed as a multi-dimensional paraboloid (χ^2, *top*), and as a multi-dimensional Gaussian ($\exp(-\chi^2/2) \approx \mathcal{L}$, *bottom*).

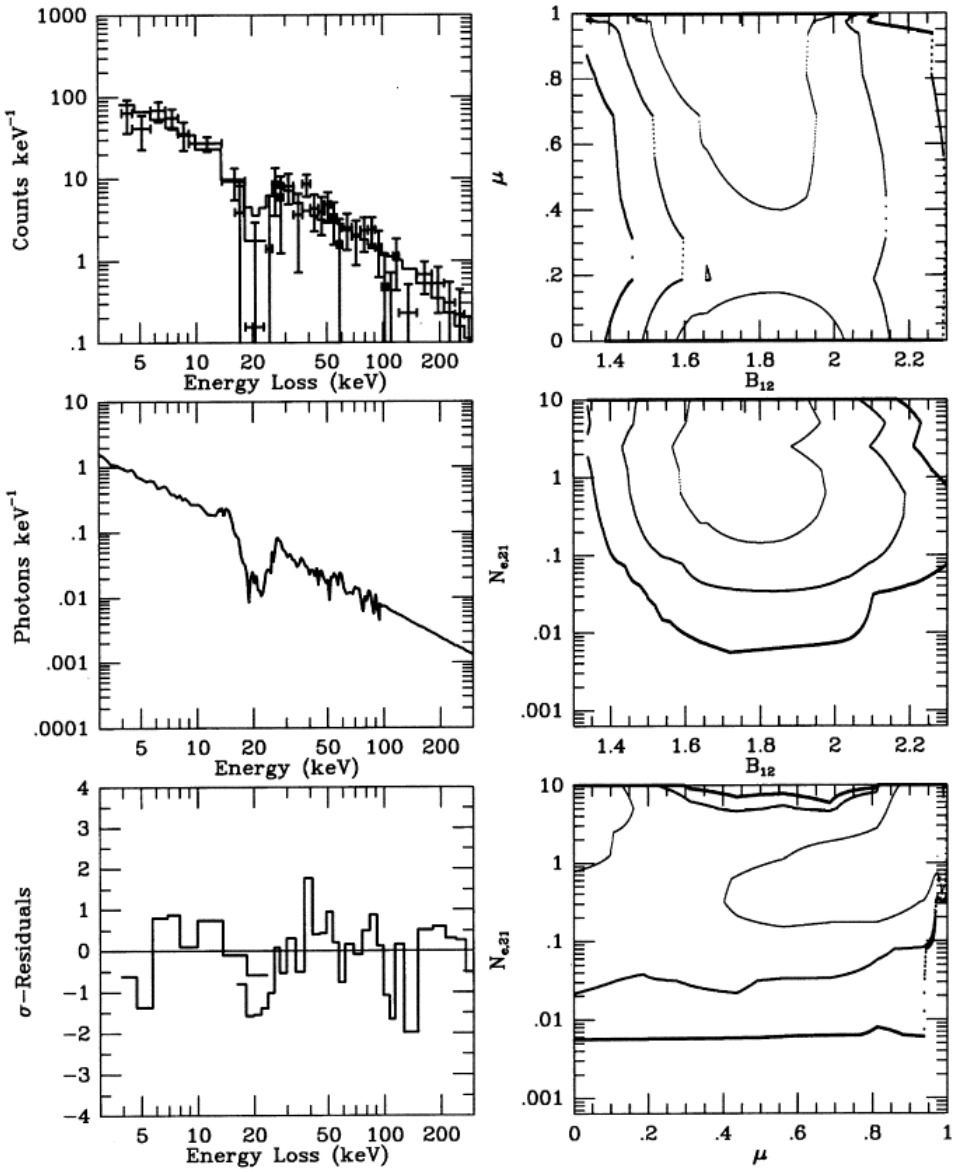


Figure 3: On the right, 1, 2, and 3σ contours determined for a statistical surface that is not "well-behaved" in parameter space. With such a surface, rigorous parameter estimation involves simulations (frequentist approach) or numerical integration of the surface (Bayesian approach). From Freeman *et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

Frequentist Parameter Estimation

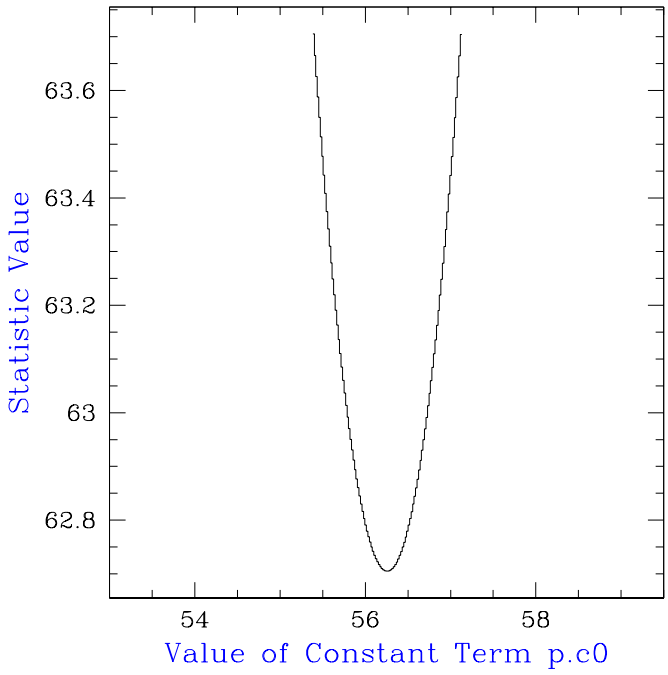
Things to keep in mind about these confidence interval estimators (dubbed `UNCERTAINTY`, `PROJECTION`, and `COVARIANCE` in *Sherpa*, respectively):

- 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.

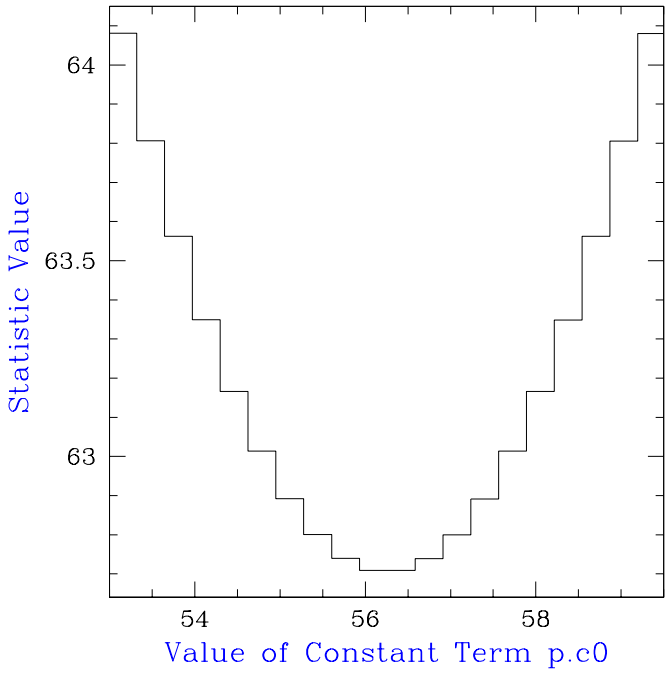
Example with a Well-Behaved Parameter Space

```
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
```

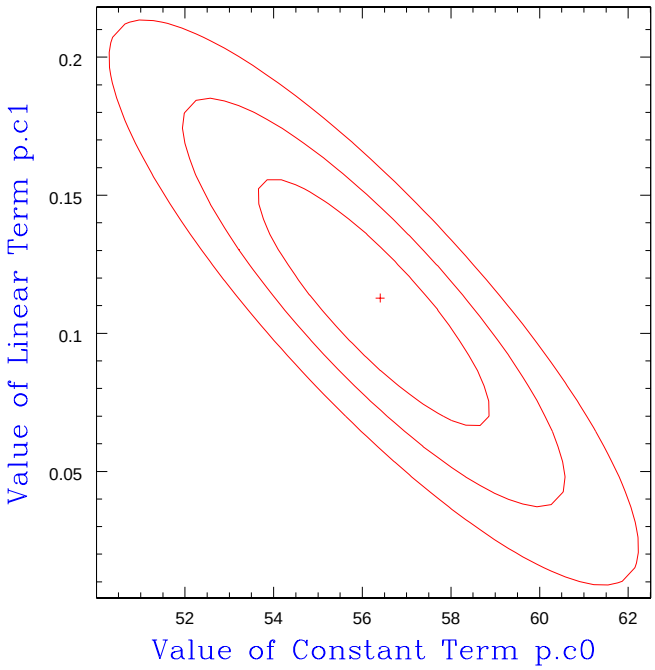
Interval – Uncertainty



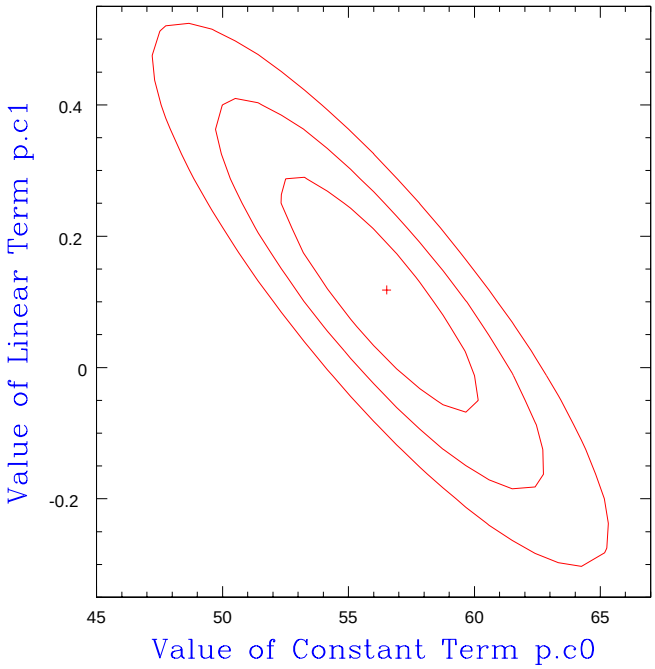
Interval – Projection



Confidence Region – Uncertainty



Confidence Region – Projection



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:

$$p(\theta_1|D) = \int_{\theta_2} d\theta_2 \cdots \int_{\theta_P} d\theta_P \; p(\theta|D) \, .$$

The central 68% of the distribution $p(\theta_1|D)$ 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:

$$\begin{aligned} p(\theta_1|D) \;\; = \;\; & p(\hat{\theta}_2, \cdots, \hat{\theta}_P)(2\pi)^{(P-1)/2} \times \\ & \sqrt{\det C(\theta_1, \hat{\theta}_2, \cdots, \hat{\theta}_P)} \mathcal{L}(\theta_1, \hat{\theta}_2, \cdots, \hat{\theta}_P) \, . \end{aligned}$$

If the value of parameter θ_1 is correlated with other parameter values, then when computing $p(\theta_1|D)$, the values of parameters $(\theta_2, \cdots, \theta_P)$ 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
- Churazov, E., et al. 1996, ApJ 471, 673
- Gehrels, N. 1986, ApJ 303, 336
- 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.h tml`
- 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`