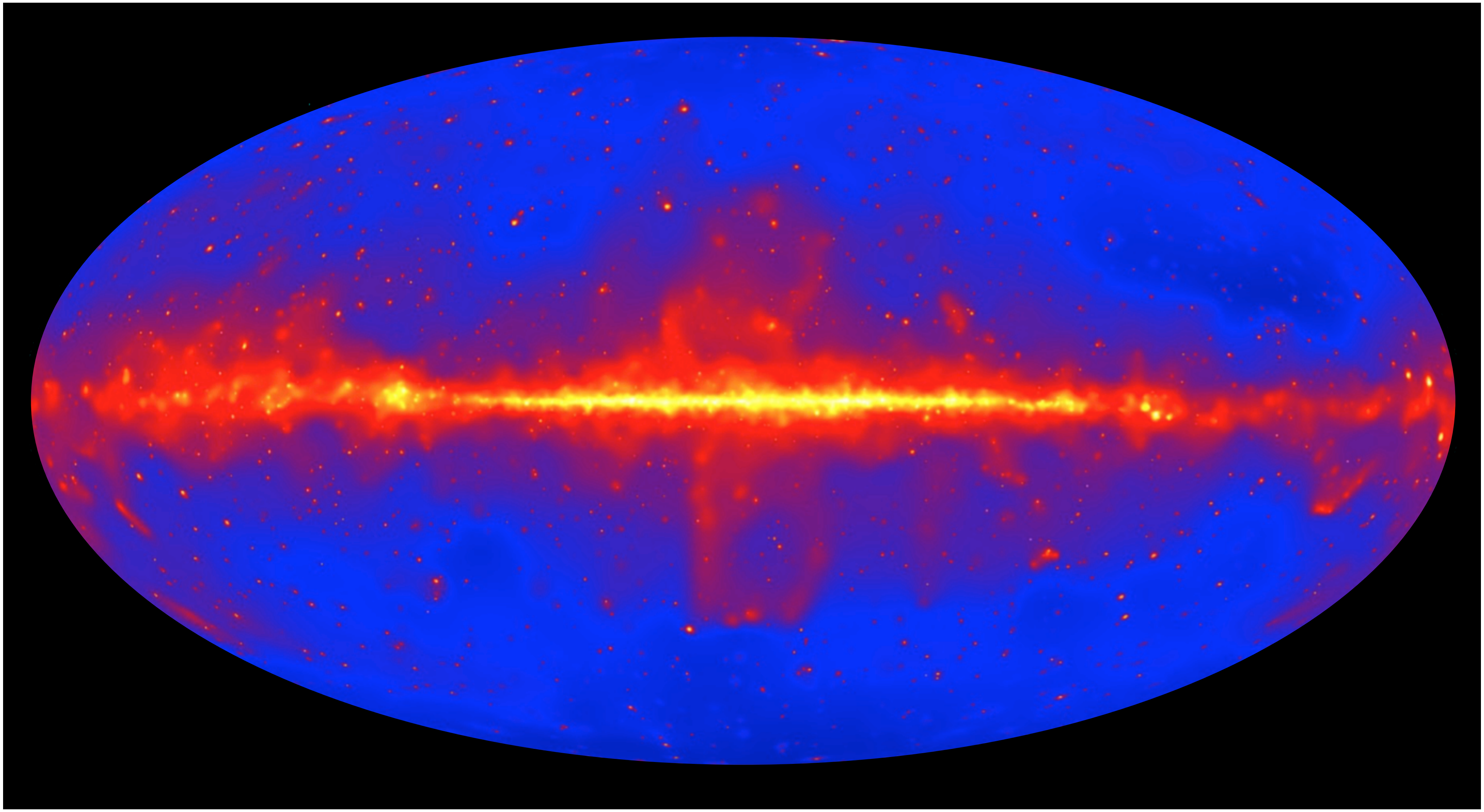
# Intro to Maximum Likelihood

Liz Hays

(heavily inspired by Steve Fegan's 2013 notes - Thanks, Steve!)

# Motivation

# Fermi LAT E>10 GeV



Fermi LAT E>10 GeV using 7 years of data (~700,000 photons)

# From LAT Event to Photon

Event → Filters

→ Reconstruction
- Direction
- Energy

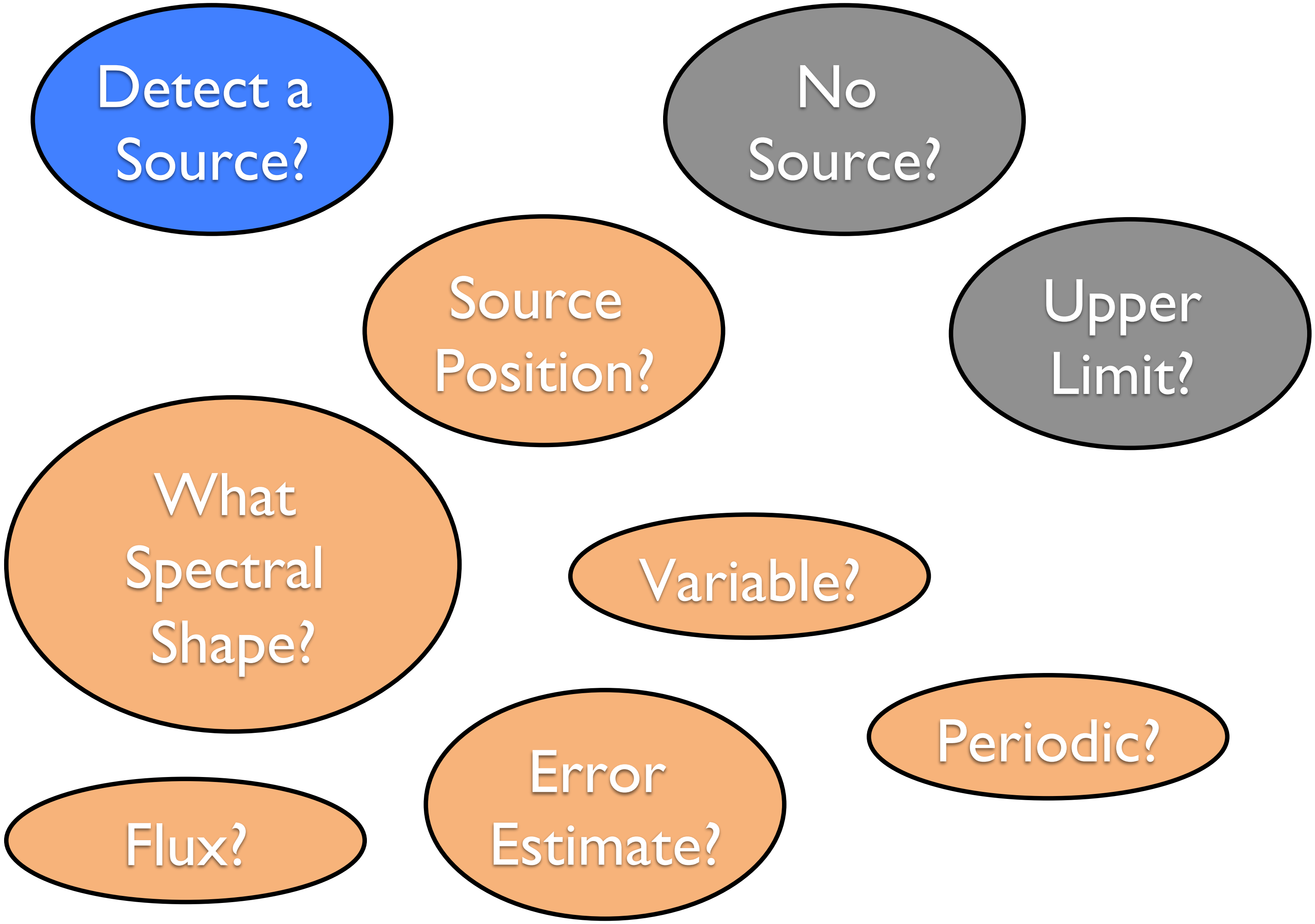
→ Classification
- Background
- Quality

→ Photon Classes
- Transient
- Source
- Clean

For each class point spread function, energy dispersion and instrumental background vary

See lecture from Jeremy

# What can I infer from my observation?



# Measurements in γ-ray astronomy

- Is a source significantly detected?
  - If so, what is its flux?
  - If not, what is upper limit on the flux?
- What kind of spectrum does it have?
  - What is its spectral index?
- What is its location in the sky?
- What are the errors on these values?
- Is the source variable?

# Measurements in γ-ray astronomy

Hypothesis testing

- Is a source significantly detected?

Parameter estimation

  - If so, what is its flux?

Parameter estimation

Hypothesis testing

  - If not, what upper limit on the flux?

Hypothesis testing

- What kind of spectrum does it have?

Parameter estimation

  - What is its spectral index?

Parameter estimation

- What is its location in the sky?

Parameter estimation

Hypothesis testing

- What are the errors on these values?

Hypothesis testing

- Is the source variable?

# The Method of Maximum Likelihood

*"a simple recipe that purports to lead to the optimum solution for all parametric problems and beyond" ~ Stigler*

Long history of evaluation: Gauss, Laplace, Fisher, Wilks…

Broad applicability to many measurement problems.

# Good things about maximum likelihood

- General framework for statistical questions.
- Unbiased, minimum variance estimate as sample size increases.
- Asymptotically Gaussian: allows evaluation of confidence bounds & hypothesis testing.
- Well studied in the literature.
- Starting point for Bayesian analysis.

# ~~Good things~~ about maximum likelihood
# Cautions

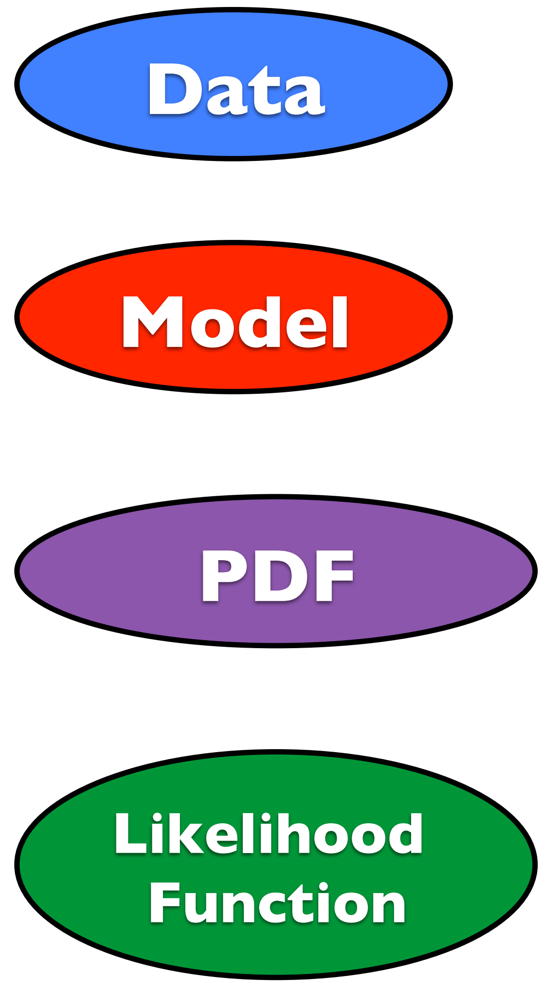
- General framework for statistical questions.
- Unbiased, minimum variance estimate as sample size increases.
- Asymptotically Gaussian: allows evaluation of confidence bounds & hypothesis testing.
- Well studied in the literature.
- Starting point for Bayesian analysis.

- Only answers the question asked.
- Be aware of small number regimes and departure from Gaussian assumption
- Starting point for Bayesian analysis.

# Maximum likelihood technique

Given a set of observed data

- produce a model that *accurately* describes the data, including parameters that we wish to estimate,
- derive the probability (density) for the data given the model (probability density function, PDF),
- treat this as a function of the model parameters (likelihood function), and
- maximize the likelihood with respect to the parameters - ML estimation.



# Maximum likelihood basics

**Data** $$X = \{x_i\} = \{x_1, x_2, ..., x_N\}$$

**Model** $$\Theta = \{\theta_j\} = \{\theta_1, \theta_2, ..., \theta_M\}$$

**Likelihood Function** $$\mathcal{L}(\Theta|X) = P(X|\Theta)$$

- Conditional probability rule for independent events: $P(A, B) \underset{\text{CPR}}{=} P(A)P(B|A) \underset{\text{Independence}}{=} P(A)P(B)$
- For independent data:

$$P(X|\Theta) = P(\{x_i\}|\Theta) = P(x_1|\Theta)P(x_2, .., x_N|\Theta) = \cdots$$

$$= P(x_1|\Theta)P(x_2|\Theta)\cdots P(x_N|\Theta) = \prod_i P(x_i|\Theta)$$

$$\mathcal{L}(\Theta|X) = \prod_i P(x_i|\Theta)$$

# ML estimation (MLE)

- Parameters can be estimated by maximizing likelihood. Easier to work with log-likelihood:

$$\boxed{\ln \mathcal{L}(\Theta) = \ln \mathcal{L}(\Theta|X) = \sum_i \ln P(x_i|\Theta)}$$

- Estimates of parameters $\{\hat{\theta}_k\}$ from solving simultaneous equations: $\left.\frac{\partial \ln \mathcal{L}}{\partial \theta_j}\right|_{\{\hat{\theta}_k\}} = 0$

- For one parameter, if we have: $\mathcal{L}(\theta) \sim e^{-\frac{(\theta-\hat{\theta})^2}{2\sigma_\theta^2}}$

  then: $\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\hat{\theta}} = -\frac{1}{\sigma_\theta^2}$

  **Gaussian approximation**

  2nd derivative is related to "errors"

# Example: $\chi^2$ fit of constant

**Data**

- independent measurements of flux of with errors $(x_i, \sigma_i)$

**Model**

- all measurements are of a constant flux with Gaussian errors $F$

**Probabilities**

$$P(x_i|F) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i - F)^2}{2\sigma_i^2}}$$

**Likelihood Function**

$$\ln \mathcal{L}(F) = -\sum \frac{(x_i - F)^2}{2\sigma_i^2} - \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi$$

# Example: $\chi^2$ fit of constant

- Log likelihood:

Constant with respect to *F*

$$\ln \mathcal{L}(F) = -\sum \frac{(x_i - F)^2}{2\sigma_i^2} \cancel{\boxed{- \sum \ln \sigma_i - \frac{N}{2} \ln 2\pi}}$$

- Maximize for MLE of $F$ :

$$\frac{\partial \ln \mathcal{L}}{\partial F} = \sum \frac{x_i - F}{\sigma_i^2} = 0 \implies \boxed{\hat{F} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}}$$

- Curvature gives "error" on *F*:

$$\frac{1}{\sigma_F^2} = -\left.\frac{\partial^2 \ln \mathcal{L}}{\partial F^2}\right|_{\hat{F}} = \sum \frac{1}{\sigma_i^2} \implies \boxed{\sigma_F = \frac{1}{\sqrt{\sum 1/\sigma_i^2}}}$$

# Example: Event counting experiment

My Gamma-ray Counter ™

**n events**

- Model: Poisson process with mean of $\lambda$:

$$P(x|\theta) \rightarrow P(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

- Log likelihood: $\ln \mathcal{L}(\lambda) = \underbrace{n \ln \lambda}_{\text{Data cpt}} - \underbrace{\lambda}_{\text{Npred}} - \overbrace{\cancel{\ln n!}}^{\text{Constant WRT } \lambda}$

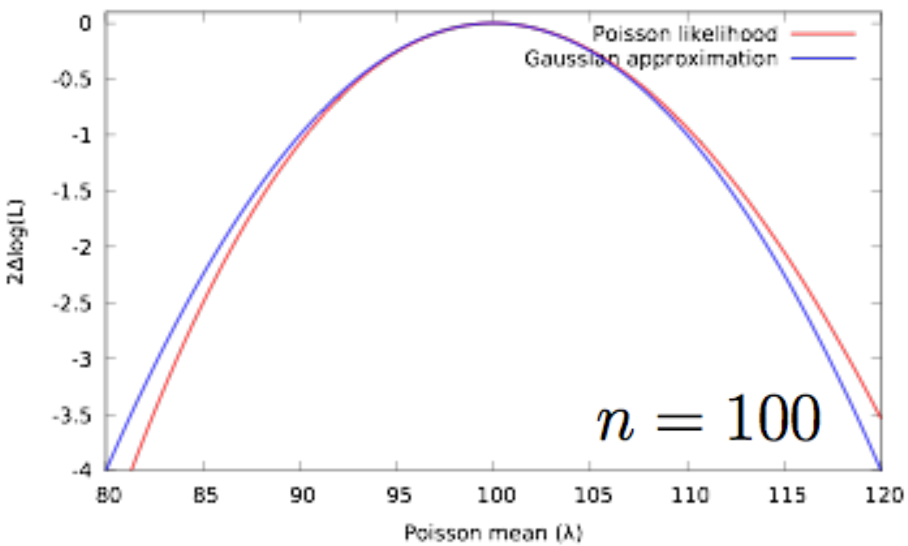
- ML estimate and error in Gaussian regime:

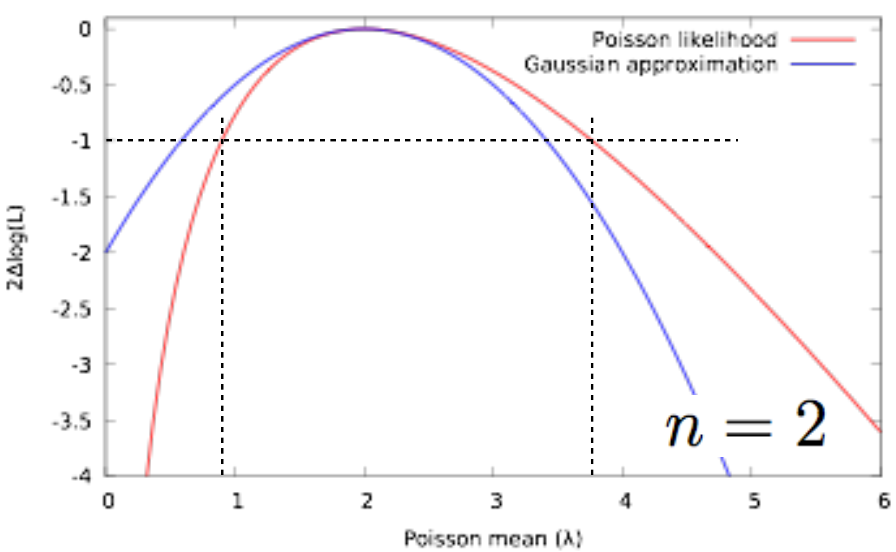
$$\frac{\partial \ln \mathcal{L}}{\partial \lambda} = \frac{n}{\lambda} - 1 \implies \boxed{\hat{\lambda} = n}$$

$$\frac{1}{\sigma_\lambda^2} = -\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \lambda^2}\right|_{\hat{\lambda}} = \frac{n}{\hat{\lambda}^2} \implies \boxed{\sigma_\lambda^2 = n}$$

**Gaussian approximation**

# Log-likelihood profile and errors




Large number of events – Gaussian approximation reasonably accurate

$$\sigma_\lambda^2 = n$$

Log-likelihood profile provides a more accurate estimate for small number of events

$$2\ln\mathcal{L}(\lambda) = 2\ln\mathcal{L}(\hat{\lambda}) - 1$$

n = 100; $\hat{\lambda} = 100.0^{+10.33}_{-9.67}$

Log-likelihood profile provides a better error estimate

n = 2; $\hat{\lambda} = 2.0^{+1.77}_{-1.10}$

# Log-likelihood profile and errors

```
# errors_poisson.py - 2013-05-07 SJF
# Evaluate the errors on the Poisson mean
import math, scipy.optimize
n_meas   = 2
logL     = lambda lam: n_meas*math.log(lam)-
lam
opt_fn   = lambda lam: -logL(lam)
opt_res  = scipy.optimize.minimize(opt_fn,
1e-8)
lam_est  = opt_res.x[0]
logL_max = logL(lam_est)
root_fn  = lambda lam: 2.0*(logL(lam)-
logL_max)+1.0
lam_lo   = scipy.optimize.brentq(root_fn,
1e-8, lam_est)
lam_hi   = scipy.optimize.brentq(root_fn,
```

# About Wilks' Theorem

- **Likelihood ratio test** compares goodness of fit of a alternate model hypothesis to a null hypothesis

- Wilks' Theorem: in limit that sample size n approaches $\infty$, the test statistic TS for nested models* is distributed like $\chi^2$ for the degrees of freedom different between the models

$$TS = 2 \ln \frac{\text{Likelihood for alternate hypothesis}}{\text{Likelihood for null hypothesis}}$$

We have a probability!

**Simulation checks highly encouraged for complicated applications*

# Confidence regions

In problems with multiple parameters.

- Saw earlier that we can calculate "asymmetric errors" by finding points where $2ln\mathcal{L}$ decreases by 1.0: 2-sided 1σ confidence interval (68%)
- Actually this comes from LRT (Wilks' theorem). This is region where null hypothesis that parameter value has some value cannot be rejected at given confidence level.
- But what to do if likelihood depends on more than our parameter of interest?
- It depends...

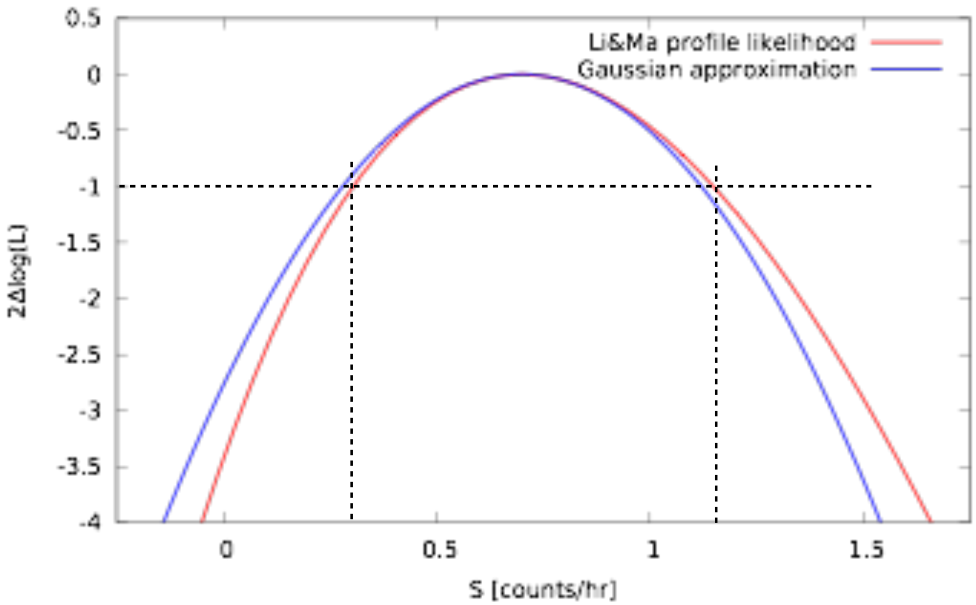
# Profile likelihood

## Confidence regions with nuisance parameters

Rolke, et al., NIM A, 551, 493 (2005)

- Often we are either concerned only with the one parameter, or wish to treat the multiple parameters separately (ignore covariance).
- Produce "profile log-likelihood" curve, a function of only one parameter (at a time), maximized over all others.
- LRT says this should behave as $\chi^2(1)$.
- Define confidence region using this function exactly as before.

# Example of profile likelihood




$$\hat{S} = 0.7^{+0.45}_{-0.39}\,\mathrm{hr}^{-1}$$

- Use simple On/Off counting example

$$n_{off} = 24$$
$$n_{on} = 15$$
$$\alpha = 1/3$$
$$T = 10.0\,\mathrm{hr}$$

- Giving:

$$\hat{S} = 0.7\,\mathrm{hr}^{-1}$$
$$\sigma_S = 0.42\,\mathrm{hr}^{-1}$$
$$TS = 3.43$$
$$\sigma = 1.85$$

This is not a significant result, so we would usually not claim a detection. Provide an upper limit instead.

# Hypothesis testing

- Compare likelihoods of two hypotheses to see which is better supported by the data.
- Likelihood-ratio test (LRT) & Wilks' theorem.
- Given a model with *N+M* parameters:

$$\Theta = \{\theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots, \theta_{N+M}\}$$

  where *N* have true values: $\theta_1^T, \ldots, \theta_N^T$

- Values of likelihood under two hypotheses:

$$\mathcal{L}_1 = \mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_N, \hat{\theta}_{N+1}, \ldots, \hat{\theta}_{N+M})$$

$$\mathcal{L}_0 = \mathcal{L}(\theta_1^T, \ldots, \theta_N^T, \hat{\theta}_{N+1}, \ldots, \hat{\theta}_{N+M})$$

- "Ratio" distributed as: $\boxed{2(\ln \mathcal{L}_1 - \ln \mathcal{L}_0) \sim \chi^2(N)}$

Terms and conditions apply

# Summary

MLE provides

- Framework for parameter estimation of a given model
- Covariant errors through inverse of Fisher matrix
- Asymmetric errors through profile likelihood
- Hypothesis testing of of models through Wilks' theorem