Introduction to Bayesian Inference
Lecture 3: Simple Examples, Curve Fitting, Questions

Tom Loredo
Dept. of Astronomy, Cornell University
http://www.astro.cornell.edu/staff/loredo/bayes/

INPE — 15 September 2009

Agenda

1. Simple examples (cont'd)
   - Normal Distribution
   - Poisson Distribution
2. Curve fitting & least-squares/χ^2
3. Question: Marginalization vs. projection
4. Question: Probability and frequency

Agenda

1. Simple examples (cont'd)
   - Normal Distribution
   - Poisson Distribution
2. Curve fitting & least-squares/χ^2
3. Question: Marginalization vs. projection
4. Question: Probability and frequency

Inference With Normals/Gaussians

Gaussian PDF

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{over } [-\infty,\infty]$$

Common abbreviated notation: $x \sim N(\mu,\sigma^2)$

Parameters

$$\begin{aligned} \mu &= \langle x \rangle \equiv \int dx\, x\, p(x|\mu,\sigma) \\ \sigma^2 &= \langle (x-\mu)^2 \rangle \equiv \int dx\, (x-\mu)^2\, p(x|\mu,\sigma) \end{aligned}$$

Gauss's Observation: Sufficiency

Suppose our data consist of N measurements, $d_i = \mu + \epsilon_i$. Suppose the noise contributions are independent, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

$$
\begin{aligned}
p(D|\mu, \sigma, M) &= \prod_i p(d_i|\mu, \sigma, M) \\
&= \prod_i p(\epsilon_i = d_i - \mu|\mu, \sigma, M) \\
&= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(d_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{\sigma^N (2\pi)^{N/2}} e^{-Q(\mu)/2\sigma^2}
\end{aligned}
$$

Find dependence of Q on μ by completing the square:

$$
\begin{aligned}
Q &= \sum_i (d_i - \mu)^2 \qquad \text{[Note: } Q/\sigma^2 = \chi^2(\mu)\text{]} \\
&= \sum_i d_i^2 + \sum_i \mu^2 - 2\sum_i d_i \mu \\
&= \left(\sum_i d_i^2\right) + N\mu^2 - 2N\mu\overline{d} \qquad \text{where } \overline{d} \equiv \frac{1}{N}\sum_i d_i \\
&= N(\mu - \overline{d})^2 + \left(\sum_i d_i^2\right) - N\overline{d}^2 \\
&= N(\mu - \overline{d})^2 + Nr^2 \quad \text{where } r^2 \equiv \frac{1}{N}\sum_i (d_i - \overline{d})^2
\end{aligned}
$$

Likelihood depends on $\{d_i\}$ only through $\overline{d}$ and r:

$$\mathcal{L}(\mu, \sigma) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)$$

The sample mean and variance are *sufficient statistics*.

This is a miraculous compression of information—the normal dist'n is highly *abnormal* in this respect!

Estimating a Normal Mean

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is known $\rightarrow I = (\sigma, M)$.

Parameter space: μ; seek $p(\mu|D, \sigma, M)$

Likelihood

$$\begin{aligned} p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \end{aligned}$$

"Uninformative" prior

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.

This prior is *improper* unless bounded.

Prior predictive/normalization

$$\begin{aligned} p(D|\sigma, M) &= \int d\mu \; C \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &= C(\sigma/\sqrt{N})\sqrt{2\pi} \end{aligned}$$

. . . minus a tiny bit from tails, using a proper prior.

Posterior

$$p(\mu|D,\sigma,M) = \frac{1}{(\sigma/\sqrt{N})\sqrt{2\pi}} \exp\left(-\frac{N(\mu-\overline{d})^2}{2\sigma^2}\right)$$

Posterior is $N(\overline{d}, w^2)$, with standard deviation $w = \sigma/\sqrt{N}$.

68.3% HPD credible region for μ is $\overline{d} \pm \sigma/\sqrt{N}$.

Note that C drops out $\rightarrow$ limit of infinite prior range is well behaved.

Informative Conjugate Prior

Use a normal prior, $\mu \sim N(\mu_0, w_0^2)$.

"Conjugate" because the posterior is also normal.

Posterior

Normal $N(\tilde{\mu}, \tilde{w}^2)$, but mean, std. deviation *"shrink"* towards prior.

Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large.

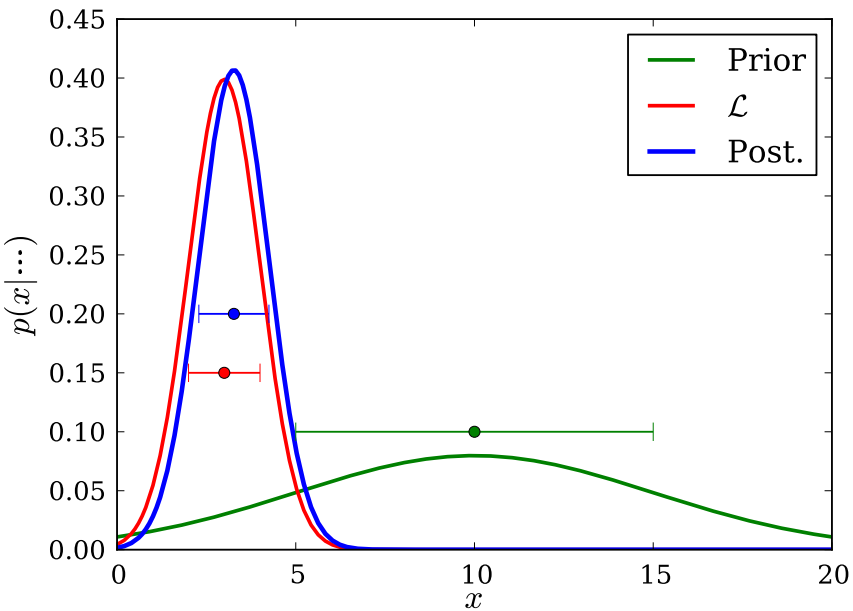
Then

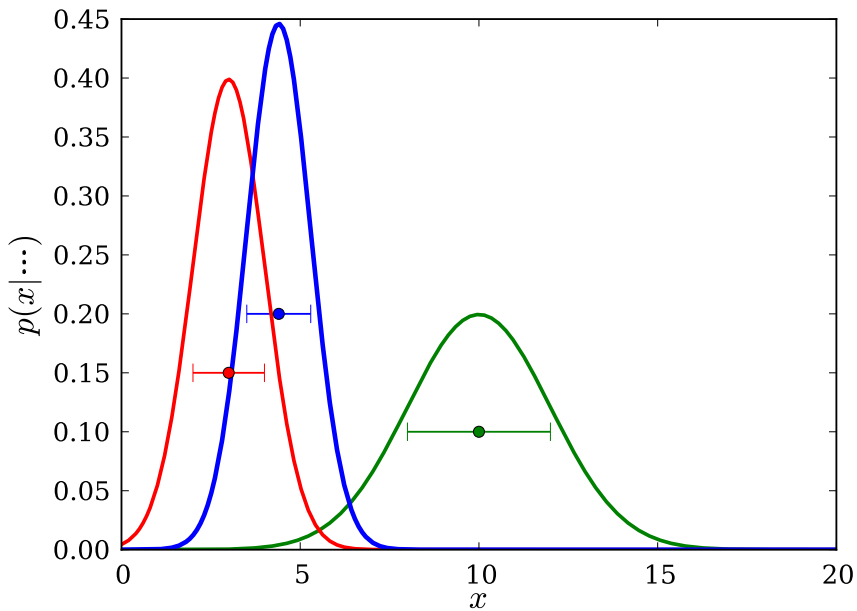
$$\begin{aligned} \widetilde{\mu} &= \overline{d} + B \cdot (\mu_0 - \overline{d}) \\ \widetilde{w} &= w \cdot \sqrt{1 - B} \end{aligned}$$

"Principle of stable estimation:" The prior affects estimates only when data are not informative relative to prior.

Conjugate normal examples:

- Data have $\overline{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$





Estimating a Normal Mean: Unknown σ

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is *unknown*

Parameter space: (μ, σ); seek $p(\mu|D, M)$

Likelihood

$$\begin{aligned} p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &\propto \frac{1}{\sigma^N} e^{-Q/2\sigma^2} \end{aligned}$$

$$\text{where} \quad Q = N\left[r^2 + (\mu - \overline{d})^2\right]$$

Uninformative Priors

Assume priors for μ and σ are independent.

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.

Scale invariance $\Rightarrow p(\sigma) \propto 1/\sigma$ (flat in $\log \sigma$).

Joint Posterior for μ, σ

$$p(\mu, \sigma | D, M) \propto \frac{1}{\sigma^{N+1}} e^{-Q(\mu)/2\sigma^2}$$

Marginal Posterior

$$p(\mu|D,M) \propto \int d\sigma \, \frac{1}{\sigma^{N+1}} e^{-Q/2\sigma^2}$$

Let $\tau = \frac{Q}{2\sigma^2}$ so $\sigma = \sqrt{\frac{Q}{2\tau}}$ and $|d\sigma| = \tau^{-3/2}\sqrt{\frac{Q}{2}}\, d\tau$

$$\begin{aligned} \Rightarrow p(\mu|D,M) &\propto 2^{N/2} Q^{-N/2} \int d\tau \, \tau^{\frac{N}{2}-1} e^{-\tau} \\ &\propto Q^{-N/2} \end{aligned}$$

Write $Q = Nr^2\left[1 + \left(\frac{\mu-\overline{d}}{r}\right)^2\right]$ and normalize:

$$p(\mu|D,M) = \frac{\left(\frac{N}{2}-1\right)!}{\left(\frac{N}{2}-\frac{3}{2}\right)!\sqrt{\pi}}\frac{1}{r}\left[1+\frac{1}{N}\left(\frac{\mu-\overline{d}}{r/\sqrt{N}}\right)^2\right]^{-N/2}$$

"Student's t distribution," with $t = \frac{(\mu-\overline{d})}{r/\sqrt{N}}$

A "bell curve," but with power-law tails

Large N:

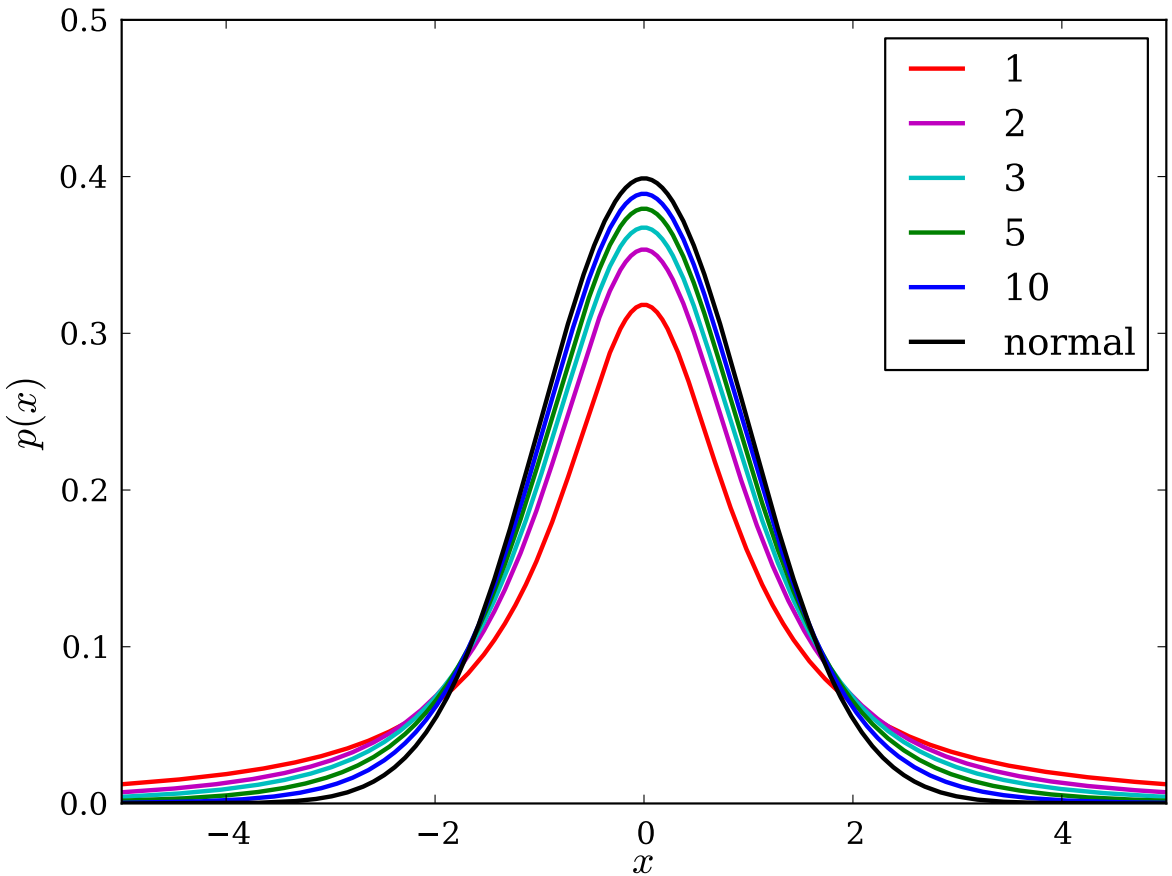
$$p(\mu|D,M) \sim e^{-N(\mu-\overline{d})^2/2r^2}$$

This is the rigorous way to "adjust σ so $\chi^2/\mathrm{dof} = 1$."

It doesn't just plug in a best σ; it slightly broadens posterior to account for σ uncertainty.

Student t examples:

- $p(x) \propto \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$
- Location $= 0$, scale $= 1$
- Degrees of freedom $= \{1, 2, 3, 5, 10, \infty\}$



Backgrounds as Nuisance Parameters

Measure background rate $b = \hat{b} \pm \sigma_b$ with source off.

Measure total rate $r = \hat{r} \pm \sigma_r$ with source on.

Infer signal source strength s, where $r = s + b$.

With flat priors,

$$p(s, b|D, M) \propto \exp\left[-\frac{(b-\hat{b})^2}{2\sigma_b^2}\right] \times \exp\left[-\frac{(s+b-\hat{r})^2}{2\sigma_r^2}\right]$$

Marginalize b to summarize the results for s (complete the square to isolate b dependence; then do a simple Gaussian integral over b):

$$p(s|D,M) \propto \exp\left[-\frac{(s-\hat{s})^2}{2\sigma_s^2}\right] \qquad \begin{array}{l} \hat{s} = \hat{r} - \hat{b} \\ \sigma_s^2 = \sigma_r^2 + \sigma_b^2 \end{array}$$

$\Rightarrow$ Background *subtraction* is a special case of background *marginalization*.

Recall the standard derivation of background uncertainty via "propagation of errors" (statistician's Delta-method) $\rightarrow$ Marginalization provides a generalization of error propagation—without approximation!

Poisson Dist'n: Infer a Rate from Counts

Problem:

Observe n counts in T; infer rate, r

Likelihood

$$\mathcal{L}(r) \equiv p(n|r,M) = p(n|r,M) = \frac{(rT)^n}{n!} e^{-rT}$$

Prior

Two simple standard choices (or conjugate gamma dist'n):

- r known to be nonzero; it is a scale parameter:

$$p(r|M) = \frac{1}{\ln(r_u/r_l)} \frac{1}{r}$$

- r may vanish; require $p(n|M) \sim$ Const:

$$p(r|M) = \frac{1}{r_u}$$

Prior predictive

$$\begin{aligned} p(n|M) &= \frac{1}{r_u}\frac{1}{n!}\int_0^{r_u} dr(rT)^n e^{-rT} \\ &= \frac{1}{r_u T}\frac{1}{n!}\int_0^{r_u T} d(rT)(rT)^n e^{-rT} \\ &\approx \frac{1}{r_u T} \quad \text{for} \quad r_u \gg \frac{n}{T} \end{aligned}$$

Posterior

A gamma distribution:

$$p(r|n, M) = \frac{T(rT)^n}{n!}e^{-rT}$$

Gamma Distributions

A 2-parameter family of distributions over nonnegative x, with shape parameter α and scale parameter s:

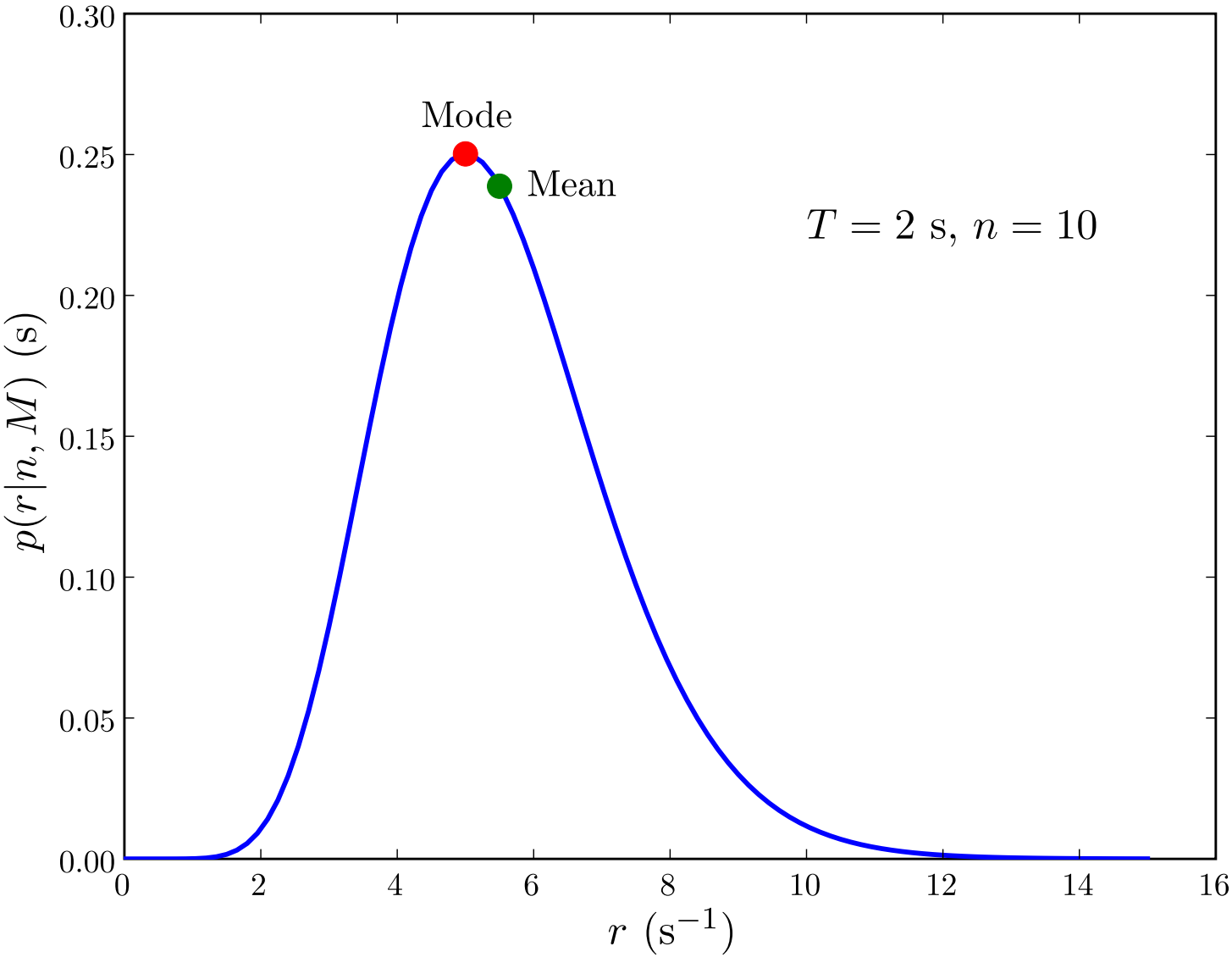
$$p_\Gamma(x|\alpha, s) = \frac{1}{s\Gamma(\alpha)} \left(\frac{x}{s}\right)^{\alpha-1} e^{-x/s}$$

Moments:

$$\mathbb{E}(x) = s\alpha \qquad \mathrm{Var}(x) = s^2\alpha$$

Our posterior corresponds to $\alpha = n + 1$, $s = 1/T$.

- Mode $\hat{r} = \frac{n}{T}$; mean $\langle r \rangle = \frac{n+1}{T}$ (shift down 1 with $1/r$ prior)
- Std. dev'n $\sigma_r = \frac{\sqrt{n+1}}{T}$; credible regions found by integrating (can use incomplete gamma function)


Mode
Mean
$T = 2$ s, $n = 10$
$p(r|n, M)$ (s)
r (s^{-1})
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
14
16

The flat prior

Bayes's justification: *Not* that ignorance of $r \rightarrow p(r|I) = C$

Require (discrete) predictive distribution to be flat:

$$\begin{aligned} p(n|I) &= \int dr\ p(r|I)p(n|r,I) = C \\ &\rightarrow \quad p(r|I) = C \end{aligned}$$

Useful conventions

- Use a flat prior for a rate that may be zero
- Use a log-flat prior ($\propto 1/r$) for a nonzero scale parameter
- Use proper (normalized, bounded) priors
- Plot posterior with abscissa that makes prior flat

The On/Off Problem

Basic problem

- Look off-source; unknown background rate b
  Count $N_{\rm off}$ photons in interval $T_{\rm off}$
- Look on-source; rate is $r = s + b$ with unknown signal s
  Count $N_{\rm on}$ photons in interval $T_{\rm on}$
- Infer s

Conventional solution

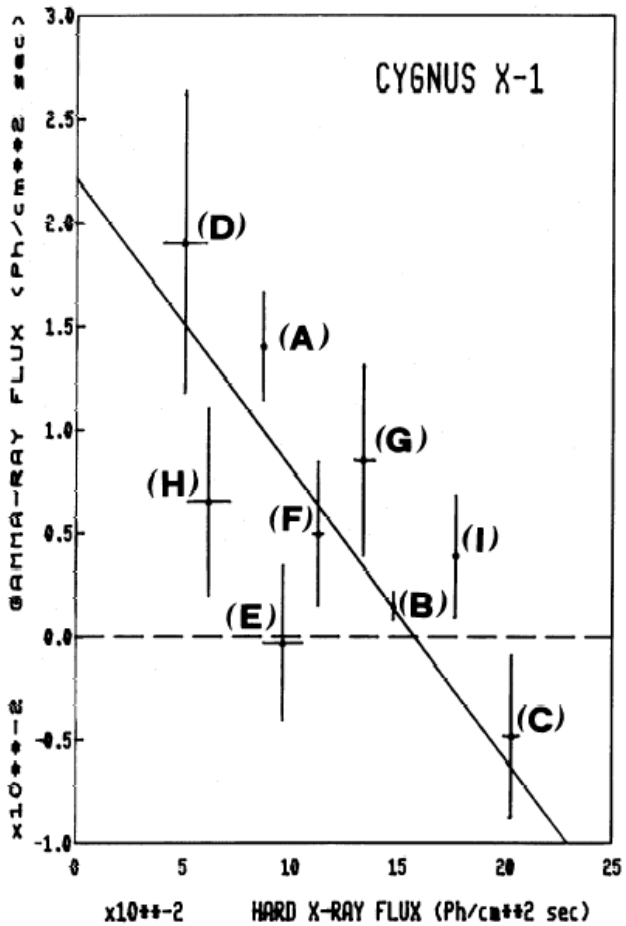
$$\begin{aligned} \hat{b} &= N_{\rm off}/T_{\rm off}; \quad & \sigma_b &= \sqrt{N_{\rm off}}/T_{\rm off} \\ \hat{r} &= N_{\rm on}/T_{\rm on}; \quad & \sigma_r &= \sqrt{N_{\rm on}}/T_{\rm on} \\ \hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2} \end{aligned}$$

But $\hat{s}$ can be negative!

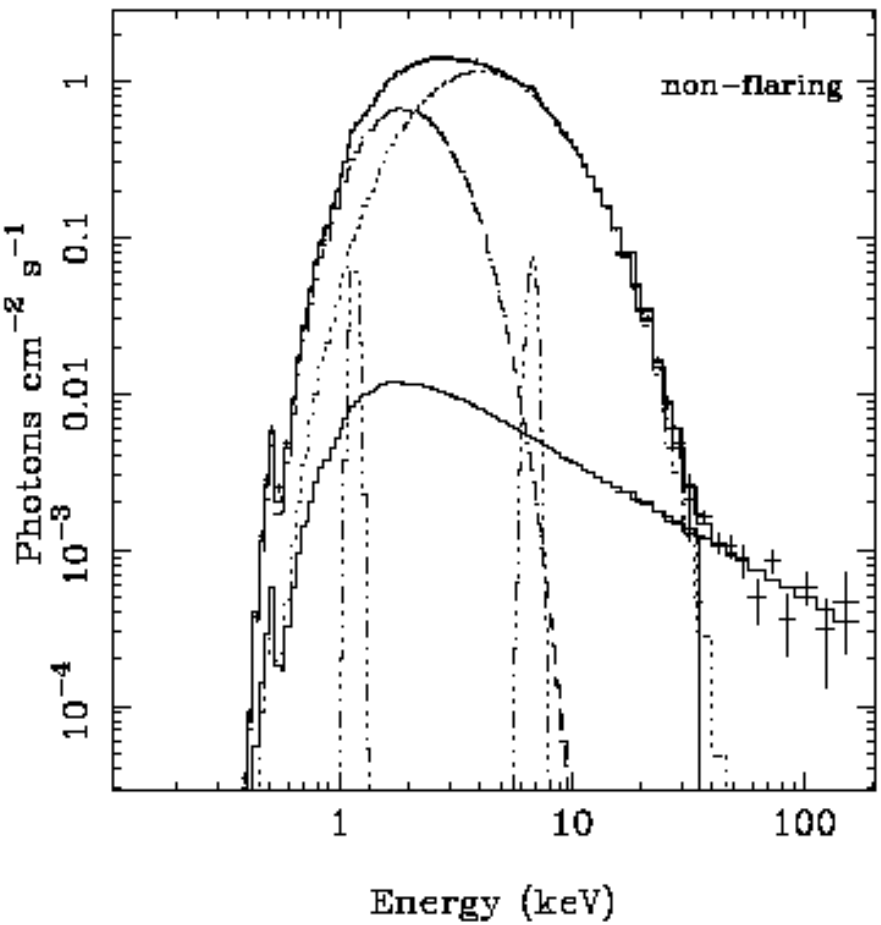
Examples

Spectra of X-Ray Sources

Bassani et al. 1989

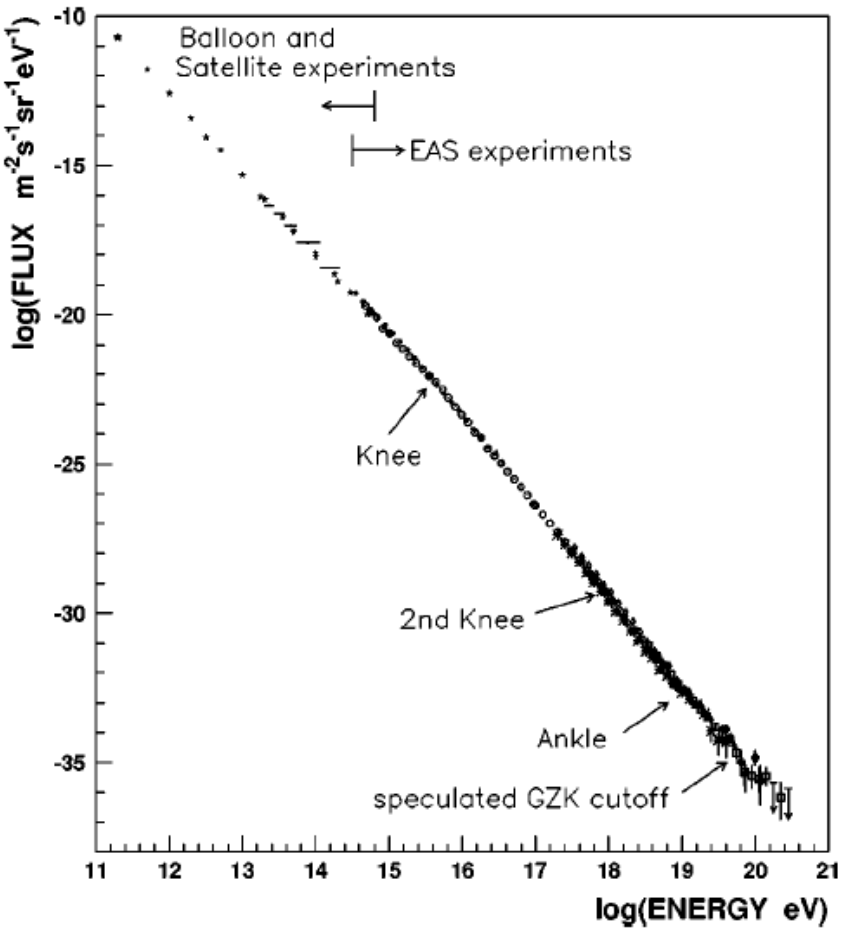


Di Salvo et al. 2001

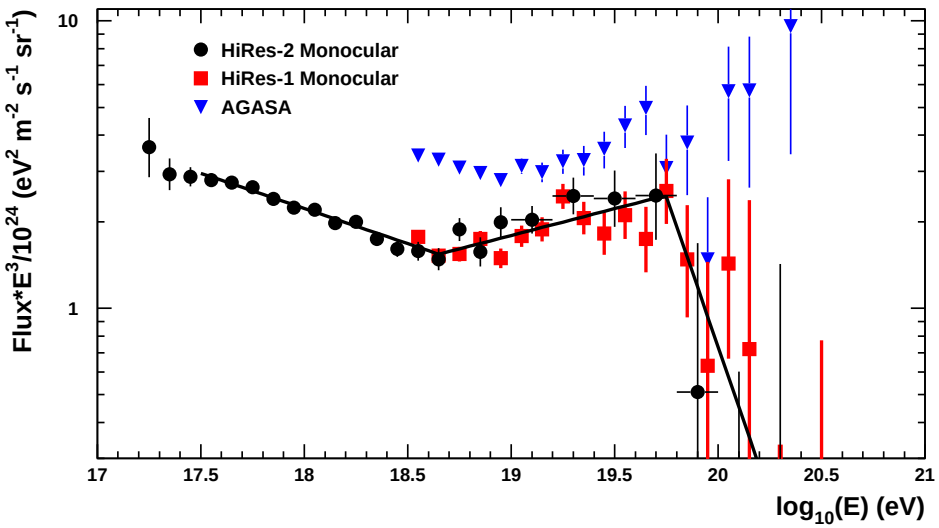


Spectrum of Ultrahigh-Energy Cosmic Rays

Nagano & Watson 2000



HiRes Team 2007



N is Never Large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

— Andrew Gelman (blog entry, 31 July 2005)

N is Never Large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

Similarly, you never have quite enough money. But that's another story.

— Andrew Gelman (blog entry, 31 July 2005)

Bayesian Solution to On/Off Problem

First consider off-source data; use it to estimate b:

$$p(b|N_{\rm off}, I_{\rm off}) = \frac{T_{\rm off}(bT_{\rm off})^{N_{\rm off}}\, e^{-bT_{\rm off}}}{N_{\rm off}!}$$

Use this as a prior for b to analyze on-source data. For on-source analysis $I_{\rm all} = (I_{\rm on}, N_{\rm off}, I_{\rm off})$:

$$p(s, b|N_{\rm on}) \quad \propto \quad p(s)p(b)[(s+b)T_{\rm on}]^{N_{\rm on}}\, e^{-(s+b)T_{\rm on}} \qquad || \; I_{\rm all}$$

$p(s|I_{\rm all})$ is flat, but $p(b|I_{\rm all}) = p(b|N_{\rm off}, I_{\rm off})$, so

$$p(s, b|N_{\rm on}, I_{\rm all}) \quad \propto \quad (s+b)^{N_{\rm on}} b^{N_{\rm off}}\, e^{-sT_{\rm on}}\, e^{-b(T_{\rm on}+T_{\rm off})}$$

Now marginalize over b;

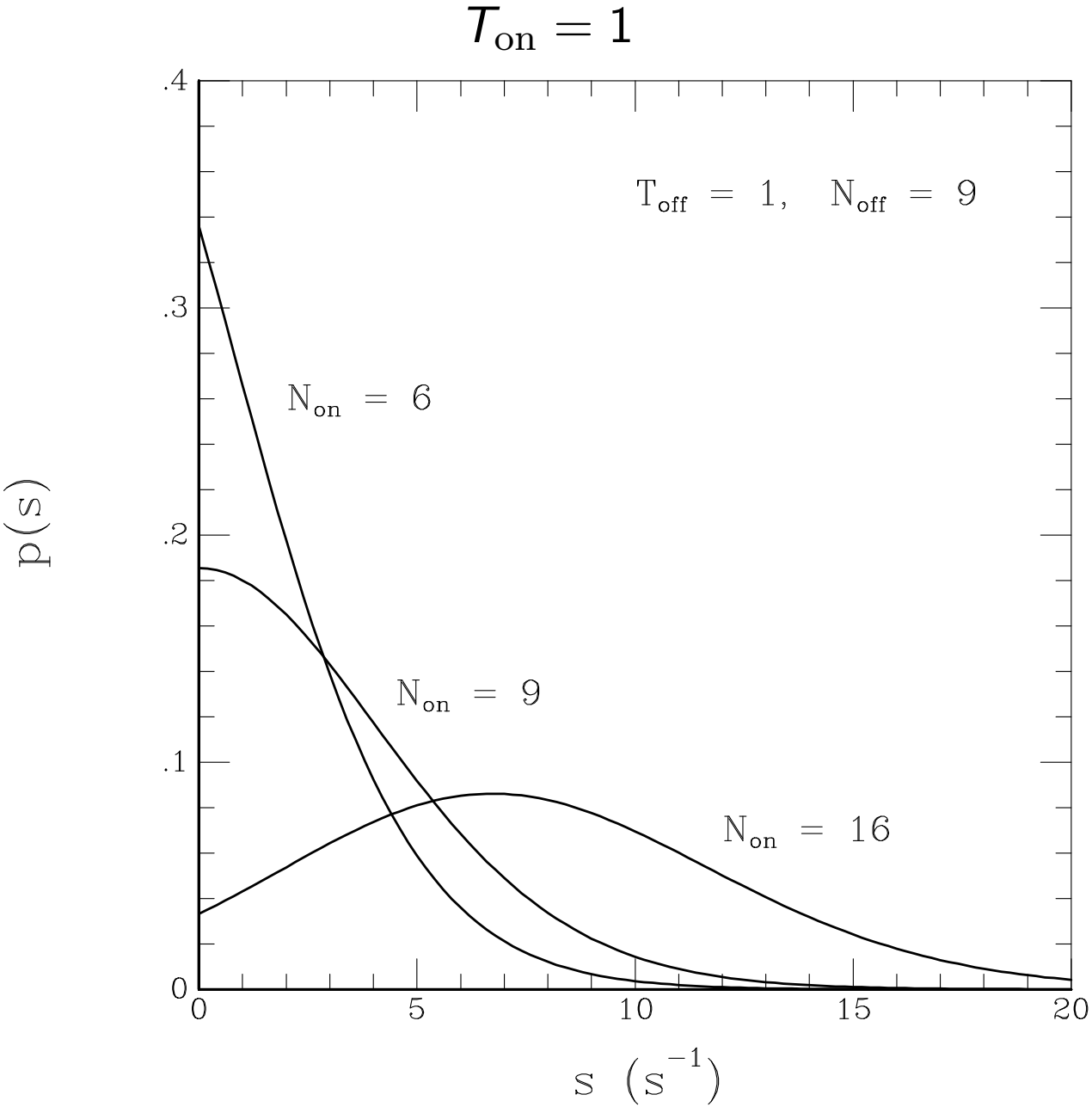
$$\begin{aligned} p(s|N_{\rm on}, I_{\rm all}) &= \int db\; p(s, b \mid N_{\rm on}, I_{\rm all}) \\ &\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})} \end{aligned}$$

Expand $(s+b)^{N_{\rm on}}$ and do the resulting Γ integrals:

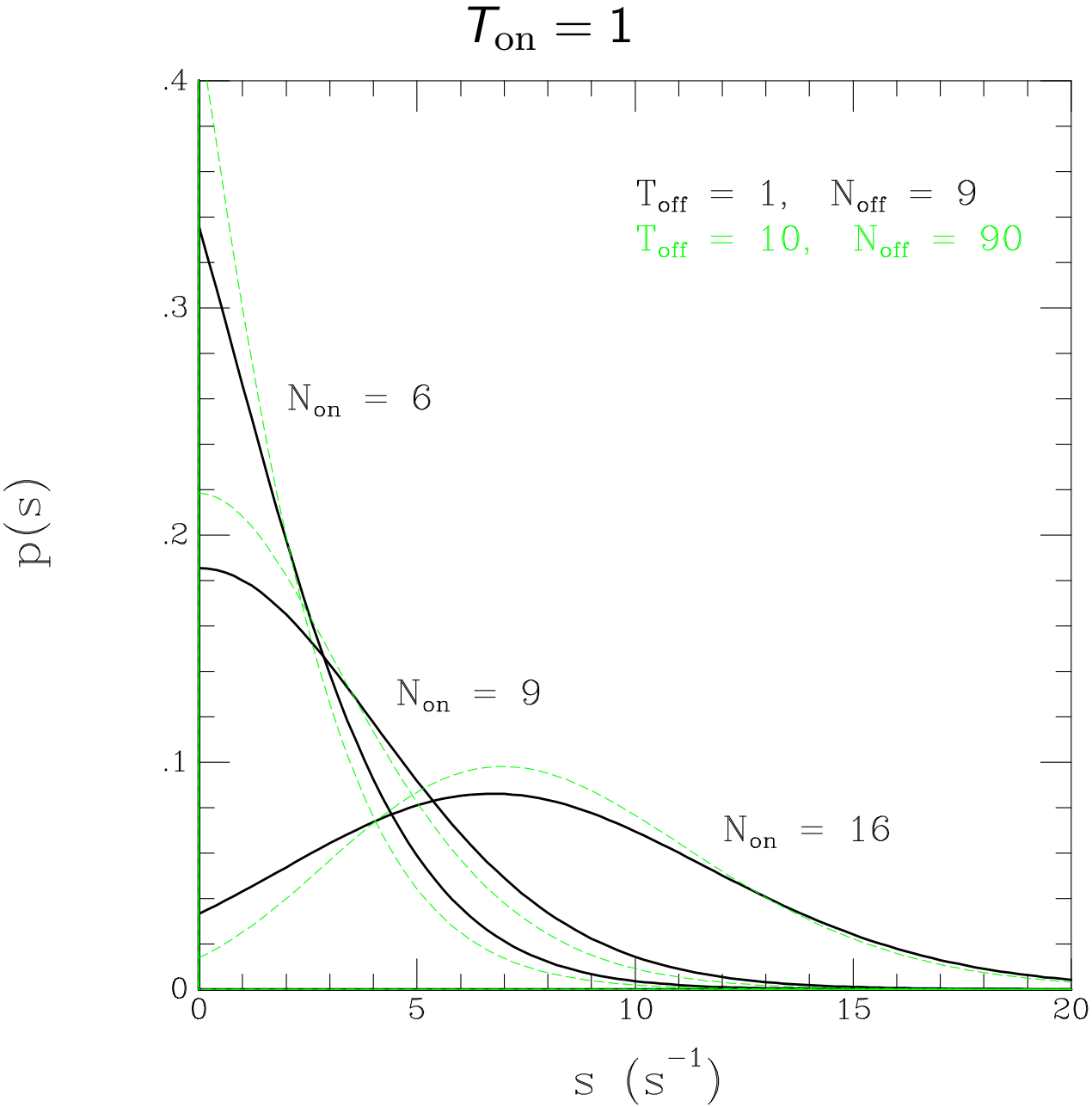
$$\begin{aligned} p(s|N_{\rm on}, I_{\rm all}) &= \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!} \\ C_i &\propto \left(1+\frac{T_{\rm off}}{T_{\rm on}}\right)^i \frac{(N_{\rm on}+N_{\rm off}-i)!}{(N_{\rm on}-i)!} \end{aligned}$$

Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)

Example On/Off Posteriors—Short Integrations



Example On/Off Posteriors—Long Background Integrations



Recap of Key Findings From Examples

- Likelihood principle
- Sufficiency: Model-dependent summary of data
- Marginalization: Generalizes background subtraction, propagation of errors
- Student's t for handling σ uncertainty
- Exact treatment of Poisson background uncertainty (don't subtract!)

Agenda

1. Simple examples (cont'd)
   - Normal Distribution
   - Poisson Distribution
2. Curve fitting & least-squares/χ^2
3. Question: Marginalization vs. projection
4. Question: Probability and frequency

Bayesian Curve Fitting & Least Squares

Setup

Data $D = \{d_i\}$ are measurements of an underlying function $f(x;\theta)$ at N sample points $\{x_i\}$. Let $f_i(\theta) \equiv f(x_i;\theta)$:

$$d_i = f_i(\theta) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_i^2)$$

We seek learn θ, or to compare different functional forms (model choice, M).

Likelihood

$$\begin{aligned} p(D|\theta, M) &= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d_i - f_i(\theta)}{\sigma_i}\right)^2\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_i \left(\frac{d_i - f_i(\theta)}{\sigma_i}\right)^2\right] \\ &= \exp\left[-\frac{\chi^2(\theta)}{2}\right] \end{aligned}$$

Bayesian Curve Fitting & Least Squares

Posterior

For prior density $\pi(\theta)$,

$$p(\theta|D, M) \propto \pi(\theta) \exp\left[-\frac{\chi^2(\theta)}{2}\right]$$

If you have a least-squares or χ^2 code:

- Think of $\chi^2(\theta)$ as $-2 \log \mathcal{L}(\theta)$.
- Bayesian inference amounts to exploration and numerical integration of $\pi(\theta)e^{-\chi^2(\theta)/2}$.

Important Case: Separable Nonlinear Models

A (linearly) separable model has parameters $\theta = (A, \psi)$:

- Linear amplitudes $A = \{A_\alpha\}$
- Nonlinear parameters ψ

$f(x;\theta)$ is a linear superposition of M nonlinear components $g_\alpha(x;\psi)$:

$$\begin{aligned} d_i &= \sum_{\alpha=1}^{M} A_\alpha g_\alpha(x_i;\psi) + \epsilon_i \\ \text{or} \quad \vec{d} &= \sum_{\alpha} A_\alpha \vec{g}_\alpha(\psi) + \vec{\epsilon}. \end{aligned}$$

Why this is important: You can marginalize over A *analytically* $\longrightarrow$ *Bretthorst algorithm* ("Bayesian Spectrum Analysis & Param. Est'n" 1988)

Algorithm is closely related to linear least squares/diagonalization.

Agenda

1. Simple examples (cont'd)
   - Normal Distribution
   - Poisson Distribution
2. Curve fitting & least-squares/χ^2
3. Question: Marginalization vs. projection
4. Question: Probability and frequency

Marginalization vs. Projection

Marginal distribution for signal s, eliminating backgrond b:

$$p(s|D,M) \quad \propto \quad p(s|M)\mathcal{L}_m(s)$$

with $\mathcal{L}_m(s)$ the *marginal likelihood for* s,

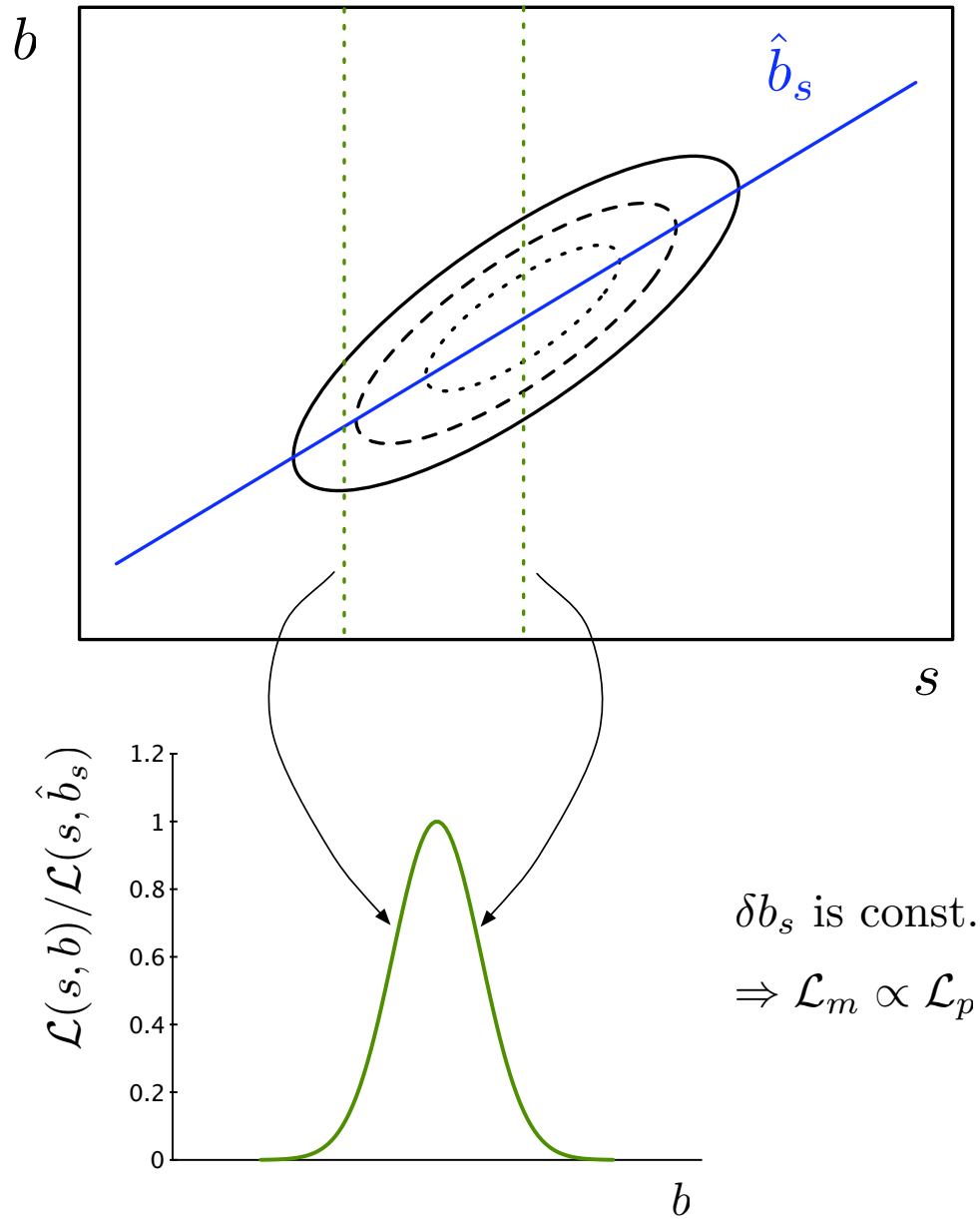
$$\begin{aligned} \mathcal{L}_m(s) &\equiv \int db\; p(b|s)\,\mathcal{L}(s,b) \\ &\approx p(\hat{b}_s|s)\;\mathcal{L}(s,\hat{b}_s)\;\delta b_s \end{aligned}$$

$\hat{b}_s$: best b given s

δb_s: b uncertainty given s

Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s,\hat{b}_s)$ gets weighted by a parameter space volume factor

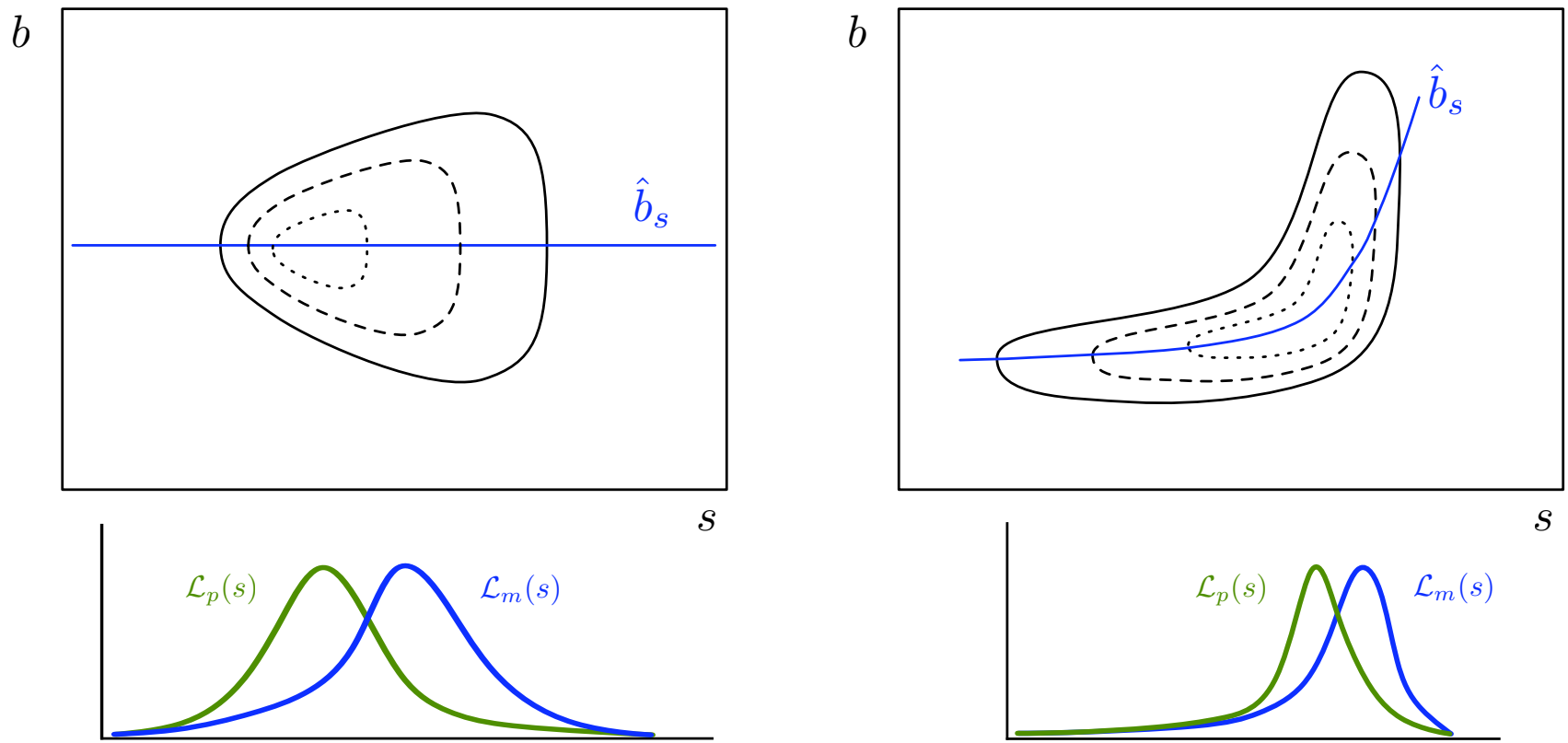
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



δb_s is const. vs. s

$\Rightarrow \mathcal{L}_m \propto \mathcal{L}_p$

Flared/skewed/bannana-shaped: $\mathcal{L}_m$ and $\mathcal{L}_p$ *differ*



General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$.
Otherwise, they will likely *differ*.

In *measurement error problems* (next lecture!) the difference becomes even greater.

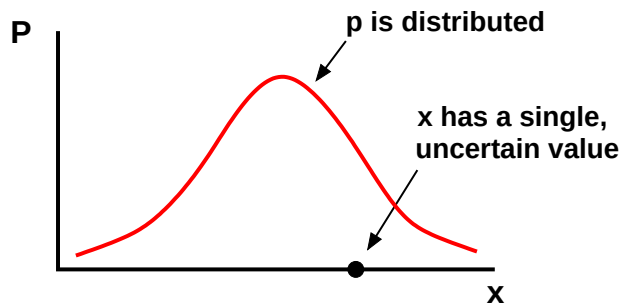
Agenda

1. Simple examples (cont'd)
   - Normal Distribution
   - Poisson Distribution
2. Curve fitting & least-squares/χ^2
3. Question: Marginalization vs. projection
4. Question: Probability and frequency

Interpreting Probability Distributions

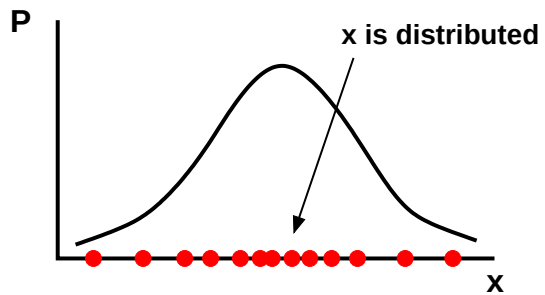
Bayesian

Probability quantifies uncertainty in an inductive inference. $p(x)$ describes how *probability* is distributed over the possible values x might have taken in the single case before us:



Frequentist

Probabilities are always (limiting) rates/proportions/frequencies in an ensemble. $p(x)$ describes variability, how the *values of* x are distributed among the cases in the ensemble:



Probability & Frequency

Frequencies are relevant when modeling repeated trials, or repeated sampling from a population or ensemble.

Frequencies are observables

- When available, can be used to *infer* probabilities for next trial
- When unavailable, can be *predicted*

Bayesian/Frequentist relationships

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures (later. . .)

Relationships Between Probability & Frequency

Frequency from probability

Bernoulli's law of large numbers: In repeated i.i.d. trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \to \alpha \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be interpreted as a frequency distribution.

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\to$ First use of Bayes's theorem:

Probability for success in next trial of i.i.d. sequence:

$$\mathbb{E}(\alpha) \to \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be estimated from a frequency distribution.

Subtle Relationships For Non-IID Cases

Predict frequency in dependent trials

r_t = result of trial t; $p(r_1, r_2 \ldots r_N|M)$ known; predict f:

$$\langle f \rangle = \frac{1}{N} \sum_t p(r_t = \text{success}|M)$$

where $$p(r_1|M) = \sum_{r_2} \cdots \sum_{r_N} p(r_1, r_2 \ldots |M_3)$$

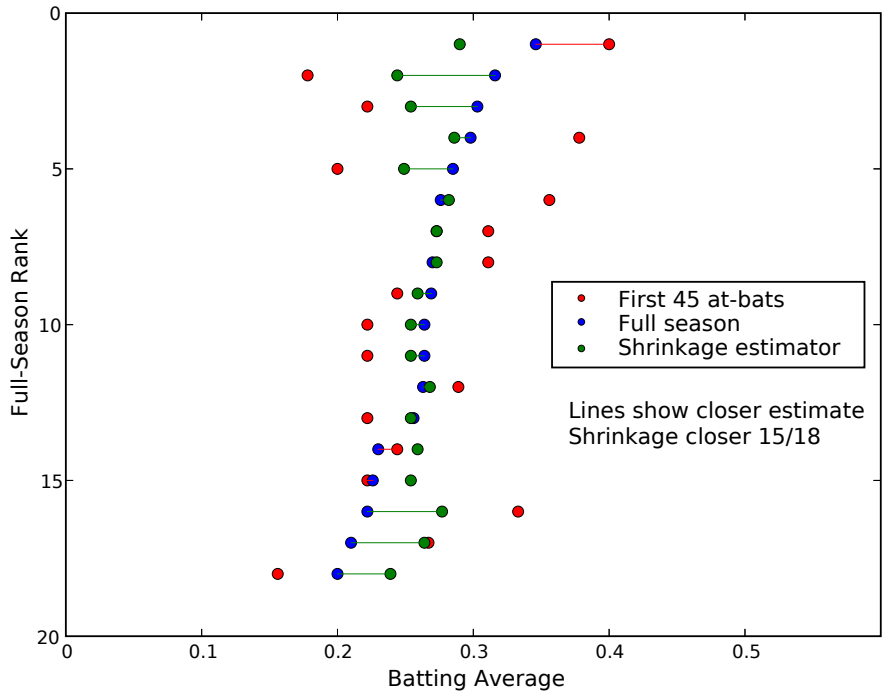
Expected frequency of outcome in many trials = *average* probability for outcome across trials.
But also find that σ_f *needn't converge to 0*.

A formalism that distinguishes p from f from the outset is particularly valuable for exploring subtle connections.

Infer probabilities for different but related trials

Batting averages for a 1970's baseball season (Efron & Morris '77):

- Red, blue: n/N (unbiased estimators)
- Green: "shrinkage estimate," n/N pulled toward overal mean



Shrinkage: Biased estimators of the probability that share info across trials are better than unbiased/BLUE/MLE estimators.

See Loredo & Hendry (2009) for connection with Eddington/Malmquist/Lutz-Kelker biases.

Frequentist Performance of Bayesian Procedures

Many results known for parametric Bayes performance:

- Estimates are consistent if the prior doesn't exclude the true value.
- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$; "reference" priors can improve their performance to $O(n^{-1})$.
- Marginal distributions have better frequentist performance than conventional methods like profile likelihood. (Bartlett correction, ancillaries, bootstrap are competitive but hard.)
- Bayesian model comparison is asymptotically consistent (not true of significance/NP tests, AIC).
- For separate (not nested) models, the posterior probability for the true model converges to 1 exponentially quickly.
- Misspecification: Bayes converges to the model with sampling dist'n closest to truth via Kullback-Leibler
- Wald's complete class theorem: *Optimal* frequentist methods are *Bayes rules* (equivalent to Bayes for some prior)
- . . .

Parametric Bayesian methods are typically good frequentist methods.

(More complicated in nonparametric problems.)