Introduction to Bayesian Inference
Lecture 2: Key Examples

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

CASt Summer School — 8 June 2012

Lecture 2: Key Examples

Key Examples

1. **Simple examples**
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
2. Multilevel models for measurement error
3. Bayesian calculation
4. Bayesian Software
5. Closing Reflections

Binary Outcomes: Parameter Estimation

M = Existence of two outcomes, S and F; for each case or trial, the probability for S is α; for F it is $(1 - \alpha)$

H_i = Statements about α, the probability for success on the next trial $\rightarrow$ seek $p(\alpha|D, M)$

D = Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood:

$$\begin{aligned} p(D|\alpha, M) &= p(\text{failure}|\alpha, M) \times p(\text{failure}|\alpha, M) \times \cdots \\ &= \alpha^n (1 - \alpha)^{N-n} \\ &= \mathcal{L}(\alpha) \end{aligned}$$

Prior

Starting with no information about α beyond its definition, use as an "uninformative" prior $p(\alpha|M) = 1$. Justifications:

- Intuition: Don't prefer any α interval to any other of same size
- Bayes's justification: "Ignorance" means that before doing the N trials, we have no preference for how many will be successes:

$$P(n\,\text{success}|M) = \frac{1}{N+1} \qquad \rightarrow \qquad p(\alpha|M) = 1$$

Consider this a *convention*—an assumption added to M to make the problem well posed.

Prior Predictive

$$\begin{aligned} p(D|M) &= \int d\alpha\, \alpha^n (1-\alpha)^{N-n} \\ &= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!} \end{aligned}$$

A *Beta integral*, $B(a,b) \equiv \int dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Posterior

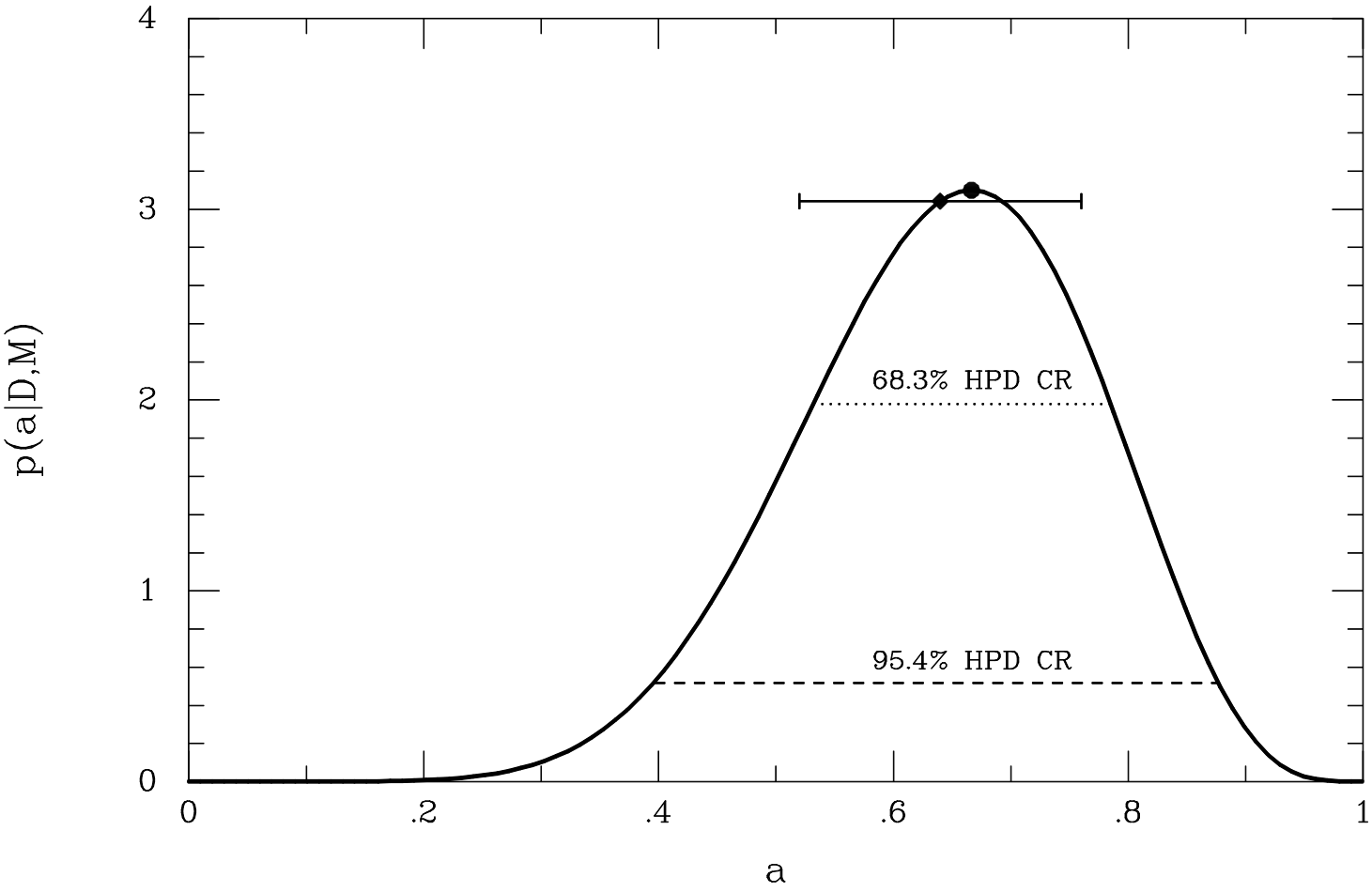
$$p(\alpha|D,M) \;=\; \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

A *Beta distribution*. Summaries:

- Best-fit: $\hat{\alpha} = \frac{n}{N} = 2/3$; $\langle\alpha\rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: $\sigma_\alpha = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$
  Find credible regions numerically, or with incomplete beta function

Note that the posterior depends on the data only through n, not the N binary numbers describing the sequence.

n is a (minimal) *sufficient statistic*.


p(a|D,M)
a
68.3% HPD CR
95.4% HPD CR
0
1
2
3
4
0
.2
.4
.6
.8
1

Binary Outcomes: Model Comparison

Equal Probabilities?

M_1: $\alpha = 1/2$
M_2: $\alpha \in [0, 1]$ with flat prior.

Maximum Likelihoods

$$M_1: \qquad p(D|M_1) = \frac{1}{2^N} = 2.44 \times 10^{-4}$$

$$M_2: \quad \mathcal{L}(\hat{\alpha}) \;=\; \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{p(D|\hat{\alpha}, M_2)} = 0.51$$

Maximum likelihoods favor M_2 (failures more probable).

Bayes Factor (ratio of model likelihoods)

$$p(D|M_1) = \frac{1}{2^N}; \qquad \text{and} \qquad p(D|M_2) = \frac{n!(N-n)!}{(N+1)!}$$

$$\begin{aligned} \rightarrow B_{12} \equiv \frac{p(D|M_1)}{p(D|M_2)} &= \frac{(N+1)!}{n!(N-n)!2^N} \\ &= 1.57 \end{aligned}$$

Bayes factor (odds) favors M_1 (equiprobable).

Note that for $n = 6$, $B_{12} = 2.93$; for this small amount of data, we can never be very sure results are equiprobable.

If $n = 0$, $B_{12} \approx 1/315$; if $n = 2$, $B_{12} \approx 1/4.8$; for extreme data, 12 flips *can* be enough to lead us to strongly suspect outcomes have different probabilities.

(Frequentist significance tests can reject null for any sample size.)

Binary Outcomes: Binomial Distribution

Suppose $D = n$ (number of heads in N trials), rather than the actual sequence. What is $p(\alpha|n, M)$?

Likelihood

Let $\mathcal{S}$ = a sequence of flips with n heads.

$$\begin{aligned} p(n|\alpha, M) &= \sum_{\mathcal{S}} p(S|\alpha, M)\, p(n|\mathcal{S}, \alpha, M) \\ &= \alpha^n (1-\alpha)^{N-n} C_{n,N} \end{aligned}$$

$p(S|\alpha, M) \to \alpha^n (1-\alpha)^{N-n}$

$p(n|\mathcal{S}, \alpha, M) \to [\![\# \text{ successes} = n]\!]$

$C_{n,N} = \#$ of sequences of length N with n heads.

$$\to p(n|\alpha, M) = \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

The *binomial distribution* for n given α, N.

Posterior

$$p(\alpha|n,M) = \frac{\frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}}{p(n|M)}$$

$$\begin{aligned} p(n|M) &= \frac{N!}{n!(N-n)!}\int d\alpha\, \alpha^n(1-\alpha)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

$$\rightarrow p(\alpha|n,M) = \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Same result as when data specified the actual sequence.

Another Variation: Negative Binomial

Suppose $D = N$, the number of trials it took to obtain a predifined number of successes, $n = 8$. What is $p(\alpha|N, M)$?

Likelihood

$p(N|\alpha, M)$ is probability for $n-1$ successes in $N-1$ trials, times probability that the final trial is a success:

$$\begin{aligned} p(N|\alpha, M) &= \frac{(N-1)!}{(n-1)!(N-n)!}\alpha^{n-1}(1-\alpha)^{N-n}\alpha \\ &= \frac{(N-1)!}{(n-1)!(N-n)!}\alpha^{n}(1-\alpha)^{N-n} \end{aligned}$$

The *negative binomial distribution* for N given α, n.

Posterior

$$p(\alpha|D,M) = C'_{n,N}\frac{\alpha^n(1-\alpha)^{N-n}}{p(D|M)}$$

$$p(D|M) \quad = \quad C'_{n,N}\int d\alpha\ \alpha^n(1-\alpha)^{N-n}$$

$$\rightarrow p(\alpha|D,M) = \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Same result as other cases.

Final Variation: "Meteorological Stopping"

Suppose $D = (N, n)$, the number of samples and number of successes in an observing run whose total number was determined by the weather at the telescope. What is $p(\alpha|D, M')$?

(M' adds info about weather to M.)

Likelihood

$p(D|\alpha, M')$ is the binomial distribution times the probability that the weather allowed N samples, $W(N)$:

$$p(D|\alpha, M') \quad = \quad W(N)\frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Let $C_{n,N} = W(N)\binom{N}{n}$. We get the *same result* as before!

Likelihood Principle

To define $\mathcal{L}(H_i) = p(D_{\text{obs}}|H_i, I)$, we must contemplate what other data we might have obtained. But the "real" sample space may be determined by many complicated, seemingly irrelevant factors; it may not be well-specified at all. Should this concern us?

Likelihood principle: The result of inferences depends only on how $p(D_{\text{obs}}|H_i, I)$ varies w.r.t. hypotheses. We can ignore aspects of the observing/sampling procedure that do not affect this dependence.

This happens because no sums of probabilities for hypothetical data appear in Bayesian results; Bayesian calculations *condition on* D_{obs}.

This is a sensible property that frequentist methods do not share. Frequentist probabilities are "long run" rates of performance, and depend on details of the sample space that are irrelevant in a Bayesian calculation.

Goodness-of-fit Violates the Likelihood Principle

Theory (H_0)

The number of "A" stars in a cluster should be 0.1 of the total.

Observations

5 A stars found out of 96 total stars observed.

Theorist's analysis

Calculate χ^2 using $\bar{n}_A = 9.6$ and $\bar{n}_X = 86.4$.

Significance level is $p(> \chi^2 | H_0) = 0.12$ (or 0.07 using more rigorous binomial tail area). Theory is accepted.

Observer's analysis

Actual observing plan was to keep observing until 5 A stars seen!

"Random" quantity is $N_{\rm tot}$, not n_A; it should follow the negative binomial dist'n. Expect $N_{\rm tot} = 50 \pm 21$.

$p(> \chi^2 | H_0) = 0.03$. Theory is rejected.

Telescope technician's analysis

A storm was coming in, so the observations would have ended whether 5 A stars had been seen or not. The proper ensemble should take into account $p(\text{storm})\ldots$

Bayesian analysis

The Bayes factor is the same for binomial or negative binomial likelihoods, and slightly favors H_0. Include $p(\text{storm})$ if you want—it will drop out!

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.

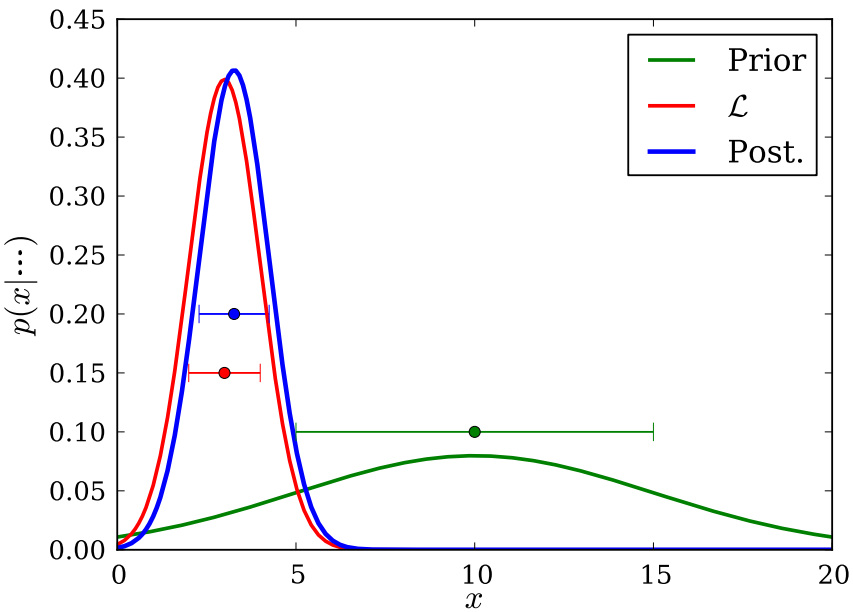
Then

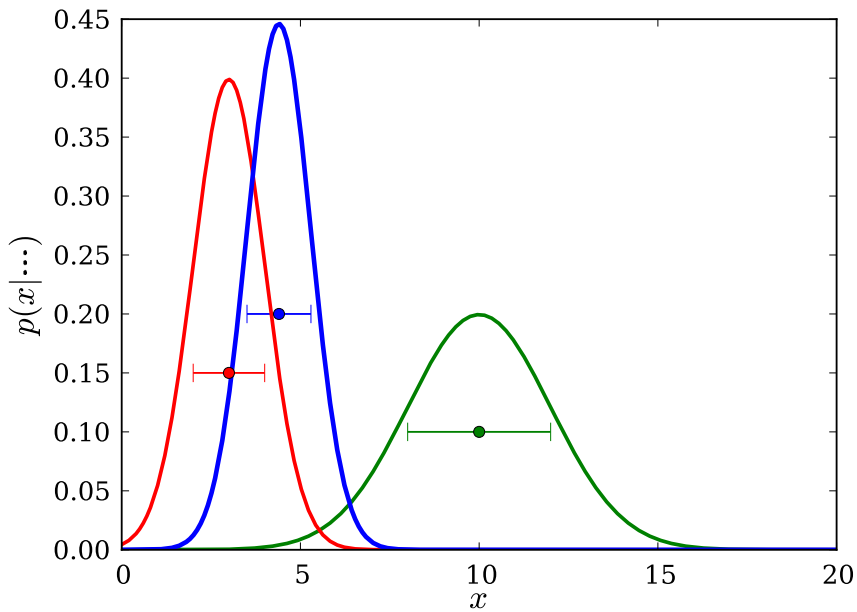
$$\begin{aligned} \widetilde{\mu} &= \overline{d} + B \cdot (\mu_0 - \overline{d}) \\ \widetilde{w} &= w \cdot \sqrt{1 - B} \end{aligned}$$

"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:

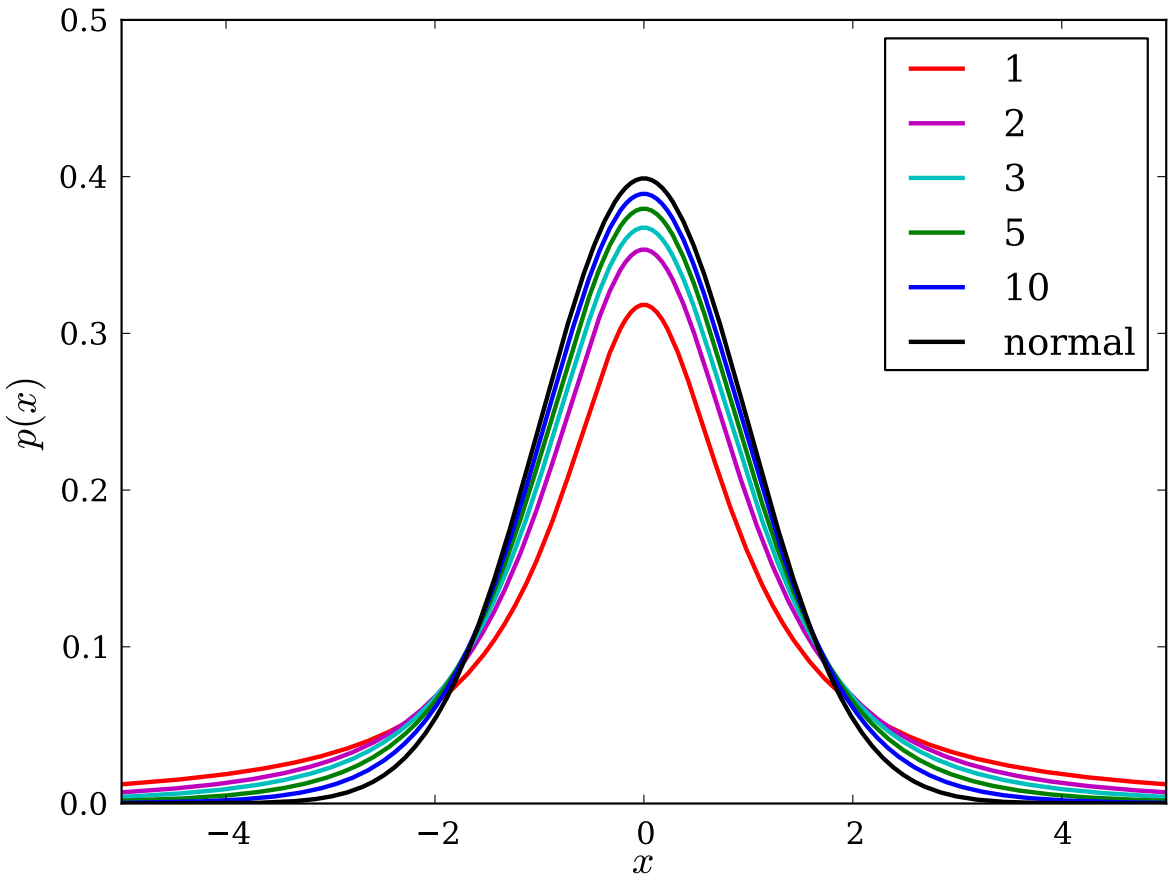
$$p(\mu|D,M) \sim e^{-N(\mu-\overline{d})^2/2r^2}$$

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 $= \{1, 2, 3, 5, 10, \infty\}$



Gaussian Background Subtraction

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*; i.e., marginalization "told us" to subtract a background estimate.

Recall the standard derivation of background uncertainty via "propagation of errors" based on Taylor expansion (statistician's *Delta-method*).

Marginalization provides a generalization of error propagation—without approximation!

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 $g_\alpha(x;\psi)$:

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

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

Algorithm is closely related to linear least squares/diagonalization/SVD.

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:

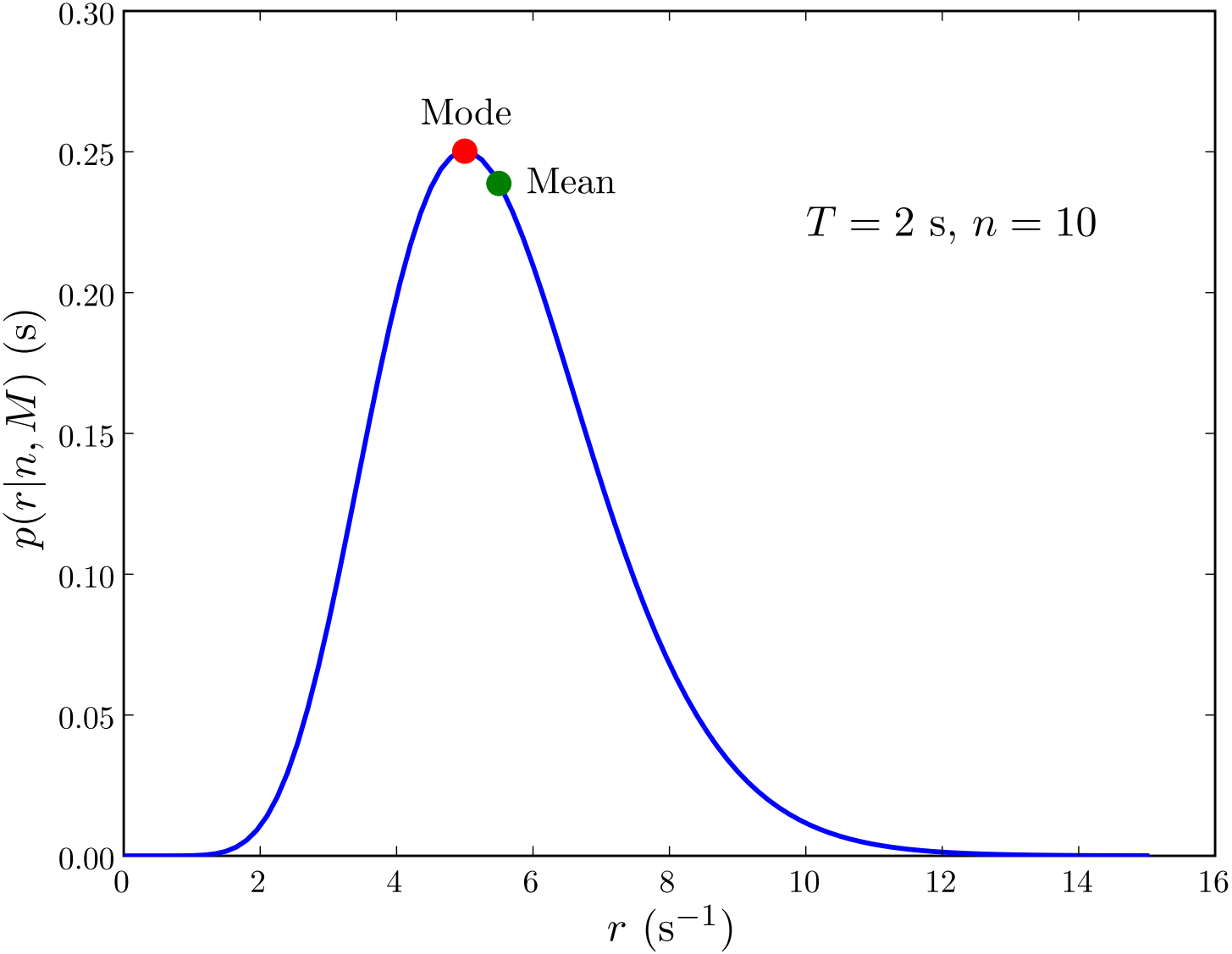
$$p_\Gamma(x|\alpha, s) = \frac{1}{s\Gamma(\alpha)} \left(\frac{x}{s}\right)^{\alpha-1} e^{-x/s}$$

Moments:

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

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})
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
12
14
16

The flat prior

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

Require (discrete) predictive distribution to be flat:

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

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

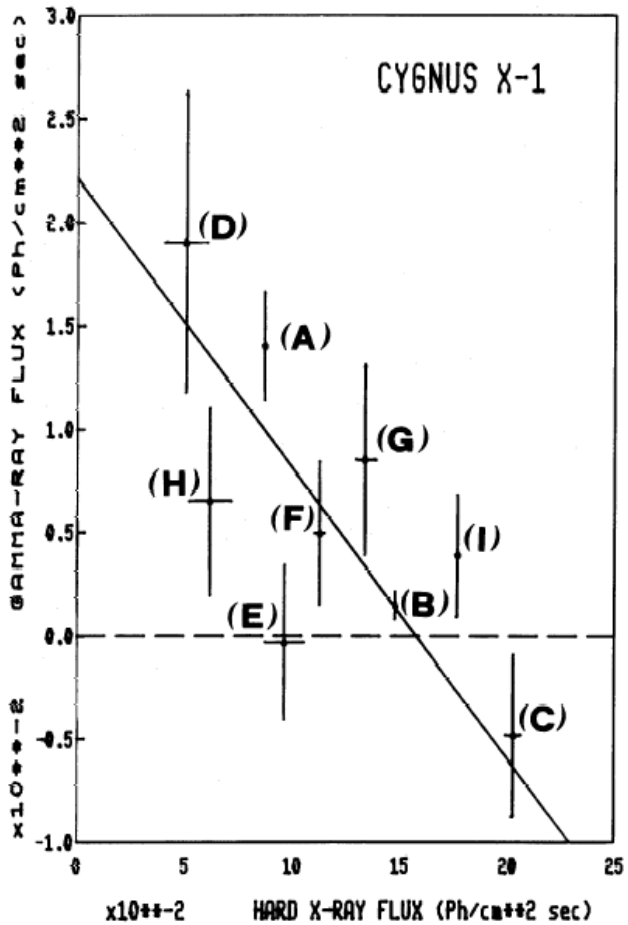
$$\begin{aligned} \hat{b} &= N_{\rm off}/T_{\rm off}; \quad & \sigma_b &= \sqrt{N_{\rm off}}/T_{\rm off} \\ \hat{r} &= N_{\rm on}/T_{\rm on}; \quad & \sigma_r &= \sqrt{N_{\rm on}}/T_{\rm on} \\ \hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2} \end{aligned}$$

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

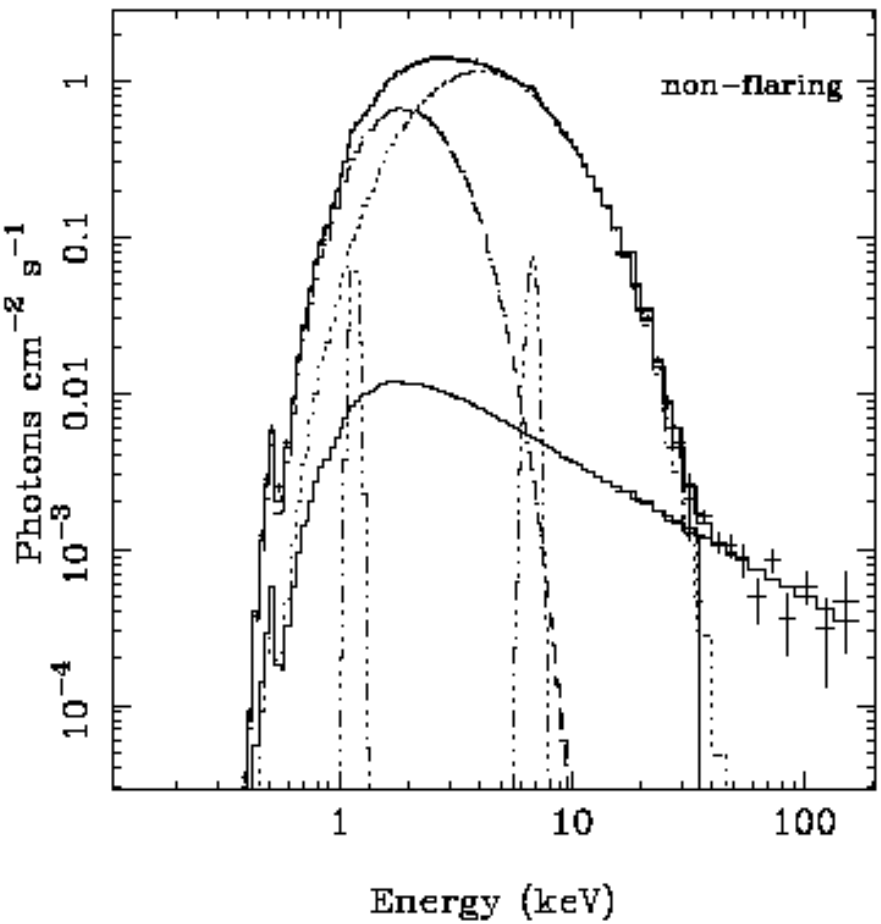
Examples

Spectra of X-Ray Sources

Bassani et al. 1989

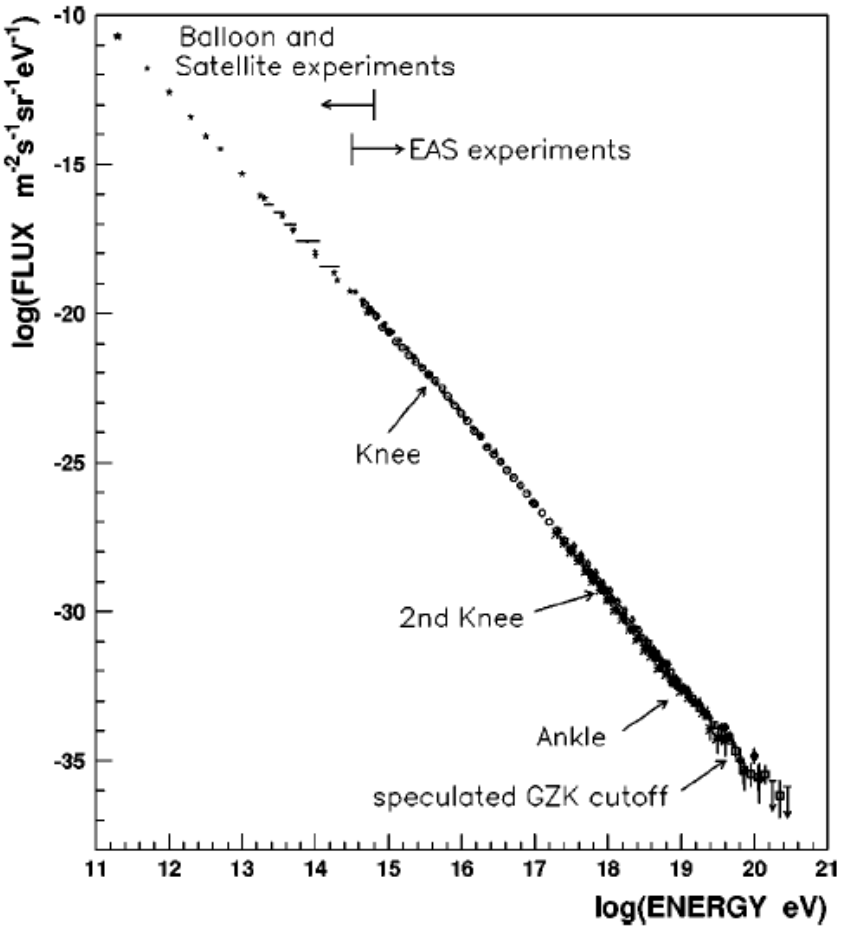


Di Salvo et al. 2001

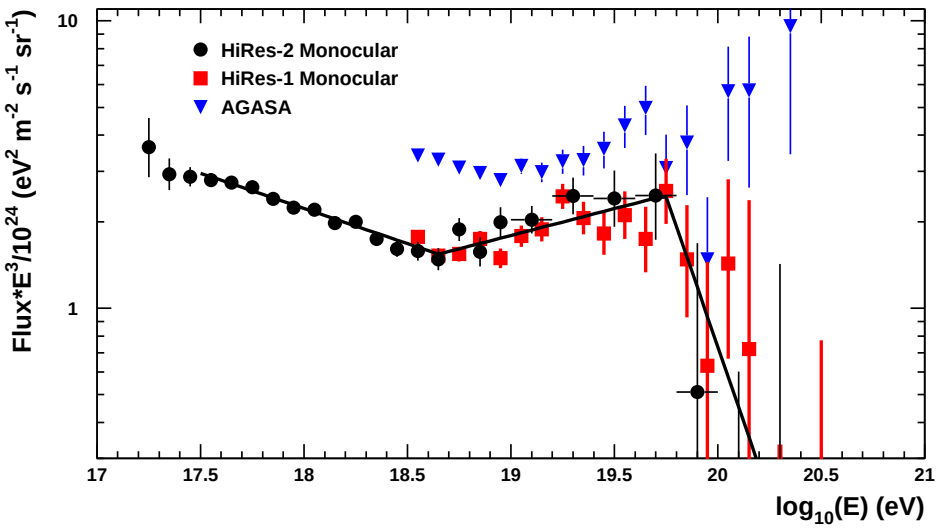


Spectrum of Ultrahigh-Energy Cosmic Rays

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;

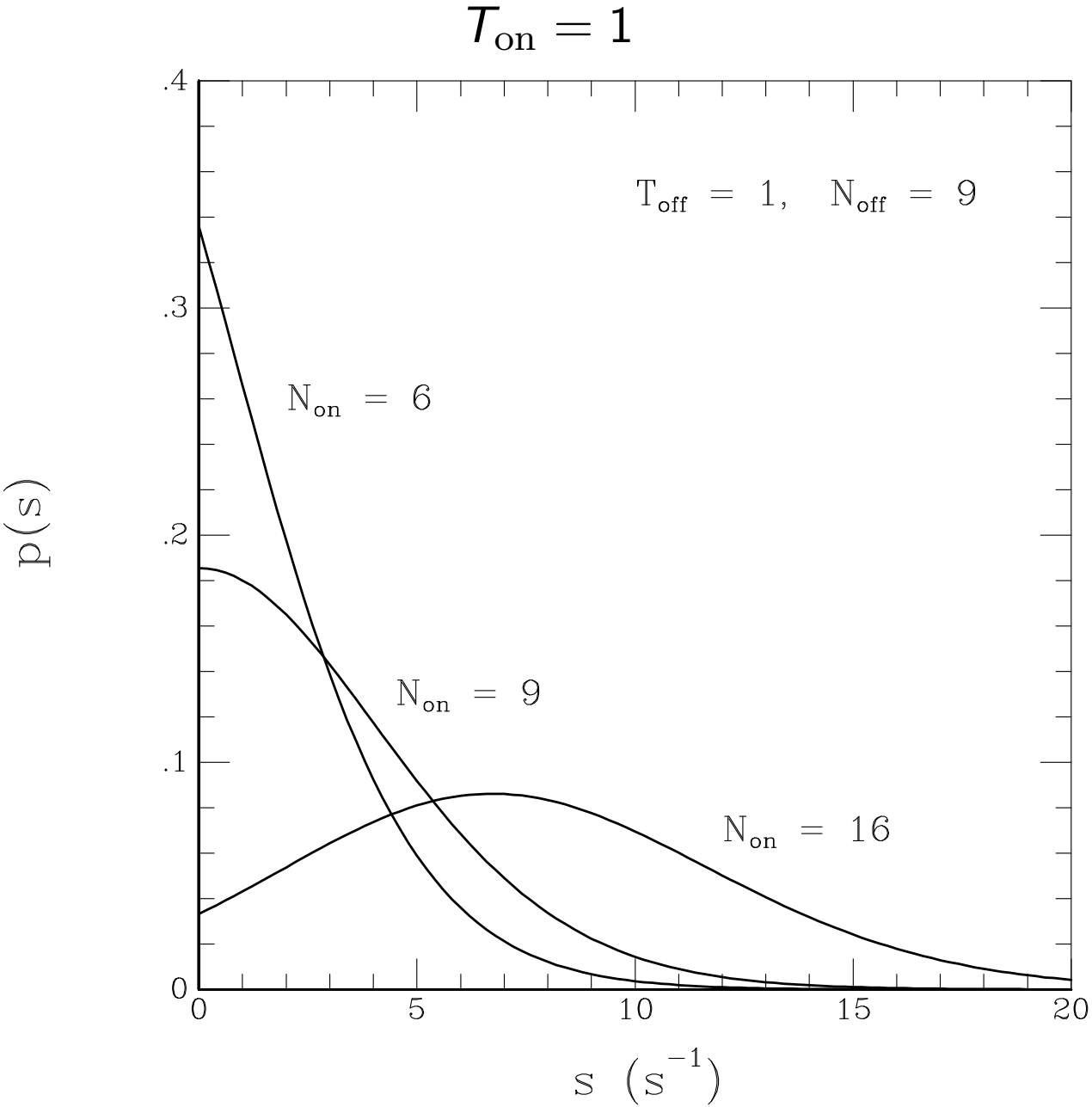
$$\begin{aligned} p(s|N_{\rm on}, I_{\rm all}) &= \int db\; p(s, b \mid N_{\rm on}, I_{\rm all}) \\ &\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})} \end{aligned}$$

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

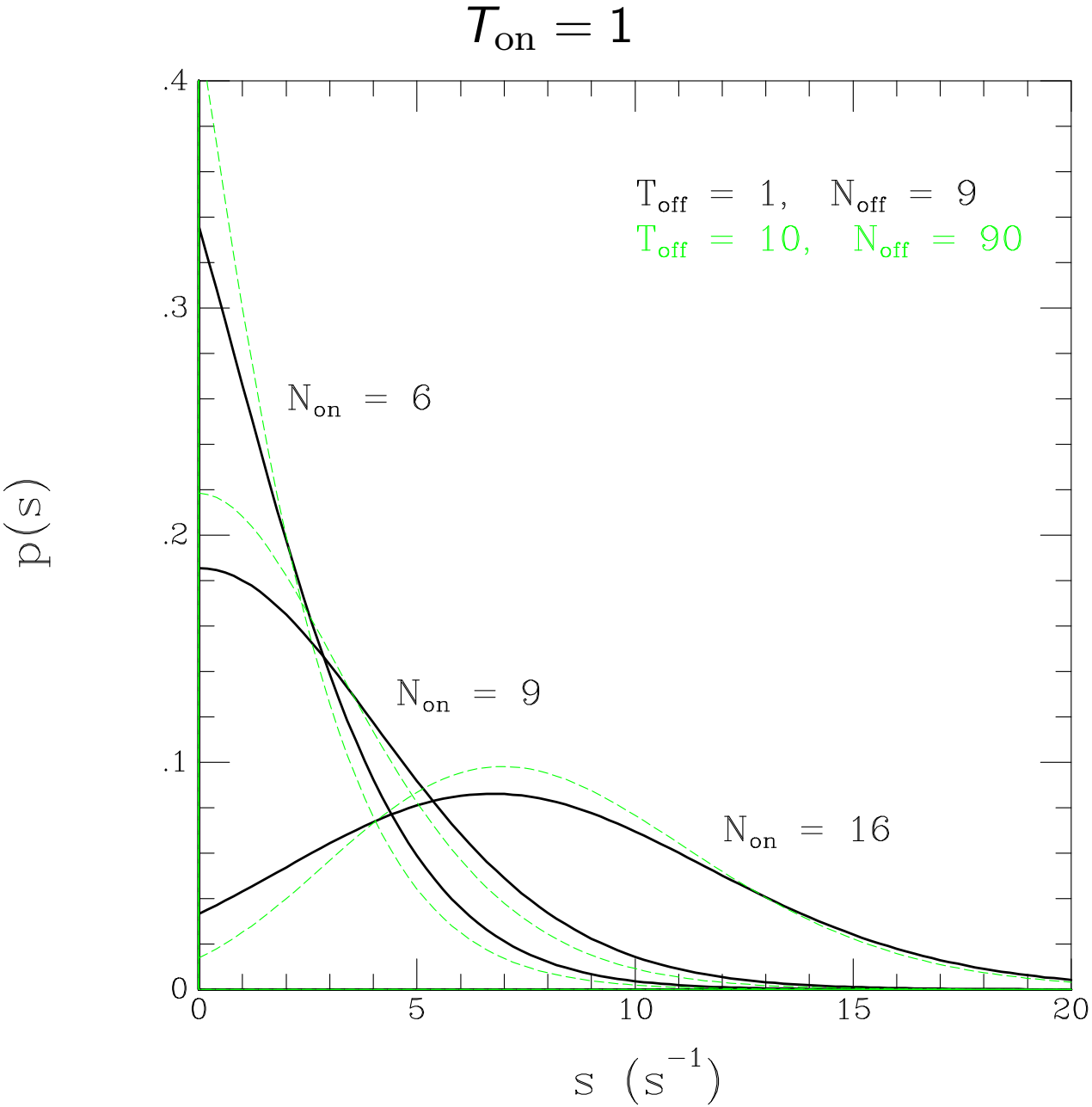
$$\begin{aligned} p(s|N_{\rm on}, I_{\rm all}) &= \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!} \\ C_i &\propto \left(1+\frac{T_{\rm off}}{T_{\rm on}}\right)^i \frac{(N_{\rm on}+N_{\rm off}-i)!}{(N_{\rm on}-i)!} \end{aligned}$$

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



Second Solution to On/Off Problem

Consider all the data at once; the likelihood is a product of Poisson distributions for the on- and off-source counts:

$$\begin{aligned}\mathcal{L}(s,b) &\equiv p(N_{\rm on}, N_{\rm off}|s,b,I) \\ &\propto [(s+b)T_{\rm on}]^{N_{\rm on}} e^{-(s+b)T_{\rm on}} \times (bT_{\rm off})^{N_{\rm off}} e^{-bT_{\rm off}}\end{aligned}$$

Take joint prior to be flat; find the joint posterior and marginalize over b;

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

$\rightarrow$ same result as before.

Third Solution: data augmentation

Suppose we knew the number of on-source counts that are from the background, N_b. Then the on-source likelihood is simple:

$$p(N_{\rm on}|s, N_b, I_{\rm all}) = {\rm Pois}(N_{\rm on} - N_b; sT_{\rm on}) = \frac{(sT_{\rm on})^{N_{\rm on}-N_b}}{(N_{\rm on} - N_b)!} e^{-sT_{\rm on}}$$

Data augmentation: Pretend you have the "missing data," then marginalize to account for its uncertainty:

$$\begin{aligned} p(N_{\rm on}|s, I_{\rm all}) &= \sum_{N_b=0}^{N_{\rm on}} p(N_b|I_{\rm all})\, p(N_{\rm on}|s, N_b, I_{\rm all}) \\ &= \sum_{N_b} \text{Predictive for } N_b \times {\rm Pois}(N_{\rm on} - N_b; sT_{\rm on}) \end{aligned}$$

$$\begin{aligned} p(N_b|I_{\rm all}) &= \int db\; p(b|N_{\rm off}, I_{\rm off})\, p(N_b|b, I_{\rm on}) \\ &= \int db\; {\rm Gamma}(b) \times {\rm Pois}(N_b; bT_{\rm on}) \end{aligned}$$

$\rightarrow$ same result as before.

A profound consistency

We solved the on/off problem in multiple ways, always finding the same final results.

This reflects something fundamental about Bayesian inference.

R. T. Cox proposed two necessary conditions for a quantification of uncertainty:

- It should duplicate deductive logic when there is no uncertainty
- Different decompositions of arguments should produce the same final quantifications (internal consistency)

Great surprise: These conditions are *sufficient*; they lead to the probability axioms. E. T. Jaynes and others refined and simplified Cox's analysis.

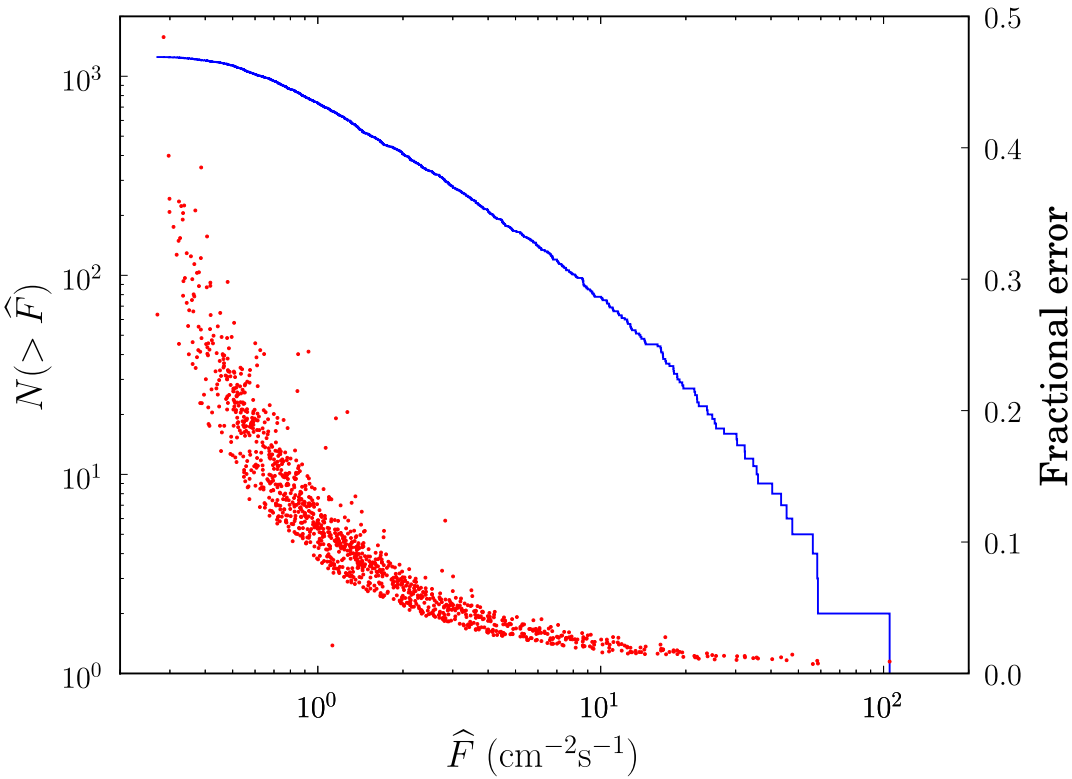
Recap of Key Findings From Examples

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

Key Examples

1. Simple examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
2. **Multilevel models for measurement error**
3. Bayesian calculation
4. Bayesian Software
5. Closing Reflections

Complications With Survey Data

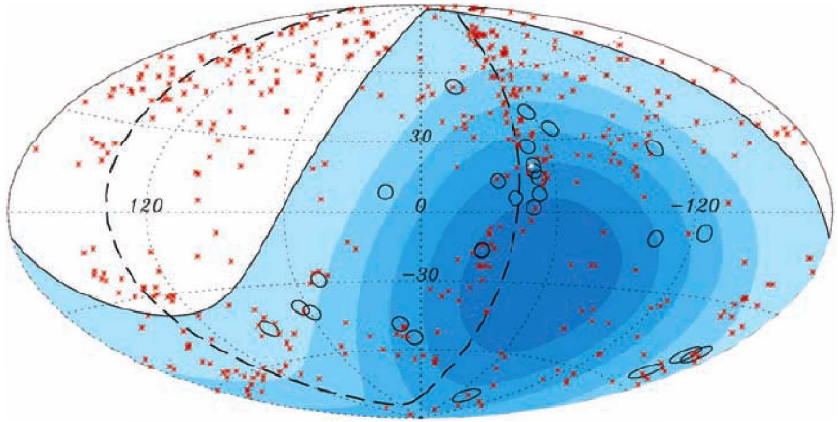


- *Selection effects* (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators

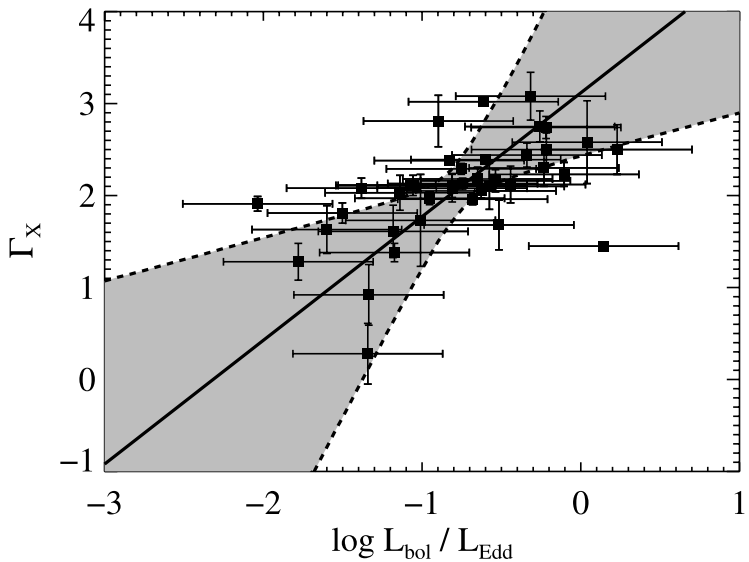
- *"Scatter" effects* (measurement error, etc.) — *insidious*
  Typically ignored (average out???)

Many Guises of Measurement Error

Auger data above GZK cutoff (PAO 2007; Soiaporn^{+} 2011)



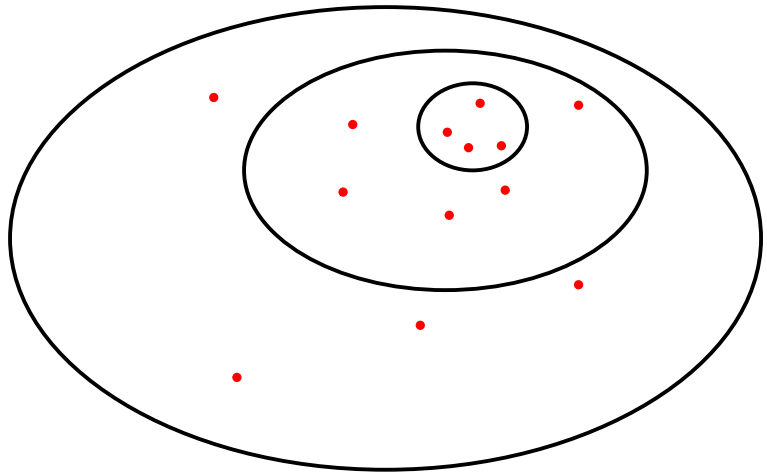
QSO hardness vs. luminosity (Kelly 2007, 2011)



Accounting For Measurement Error

Introduce latent/hidden/incidental parameters

Suppose $f(x|\theta)$ is a distribution for an observable, x.

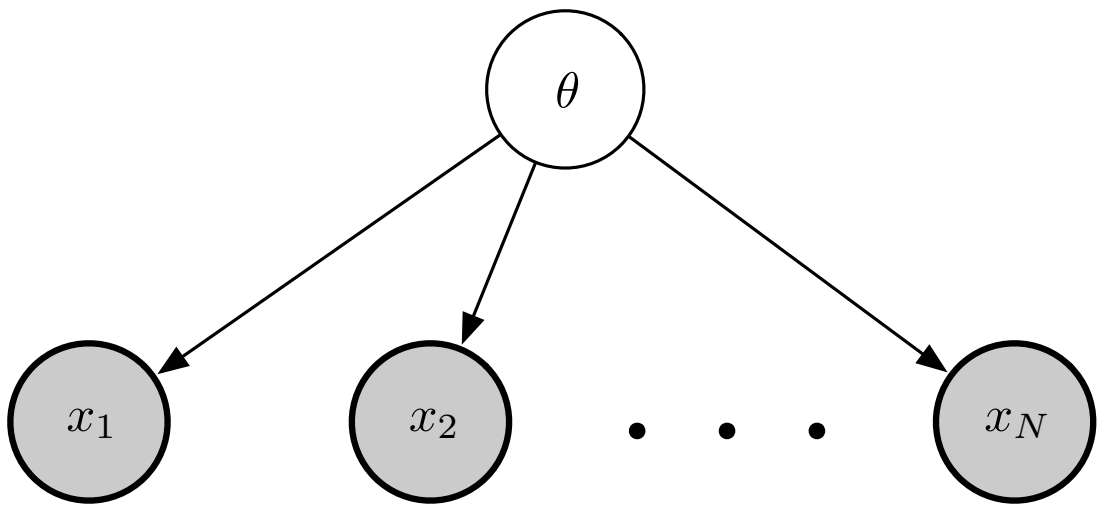


From N precisely measured samples, $\{x_i\}$, we can infer θ from

$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$

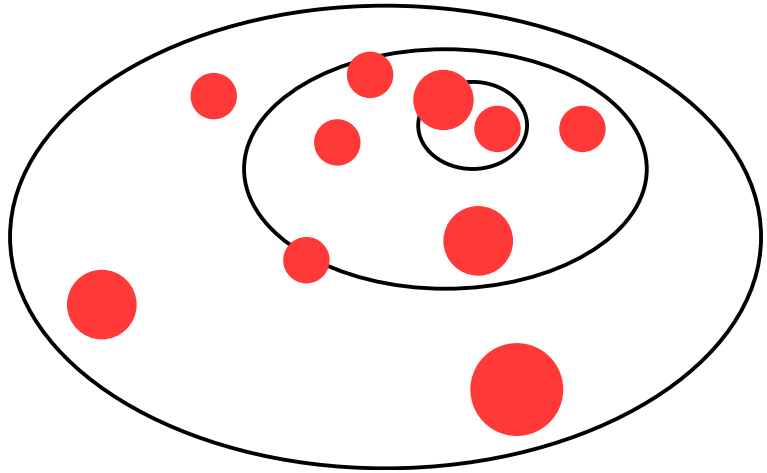
Graphical representation

- Nodes/vertices = uncertain quantities
- Edges specify conditional dependence
- Absence of an edge denotes conditional *in*dependence



$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$

But what if the x data are *noisy*, $D_i = \{x_i + \epsilon_i\}$?



We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$

$$\begin{aligned} \mathcal{L}(\theta, \{x_i\}) &\equiv p(\{D_i\}|\theta, \{x_i\}) \\ &= \prod_i \ell_i(x_i) f(x_i|\theta) \end{aligned}$$

Marginalize over $\{x_i\}$ to summarize inferences for θ.
Marginalize over θ to summarize inferences for $\{x_i\}$.

Key point: *Maximizing over x_i and integrating over x_i can give very different results!*

To estimate x_1:

$$\begin{aligned}
\mathcal{L}_m(x_1) &= \int d\theta\, \ell_1(x_1) f(x_1|\theta) \times \int dx_2 \ldots \int dx_N \prod_{i=2}^{N} \ell_i(x_i) f(x_i|\theta) \\
&= \int d\theta\, \ell_1(x_1) f(x_1|\theta) \mathcal{L}_{-1}(\theta) \\
&\approx \ell_1(x_1) f(x_1|\hat{\theta})
\end{aligned}$$

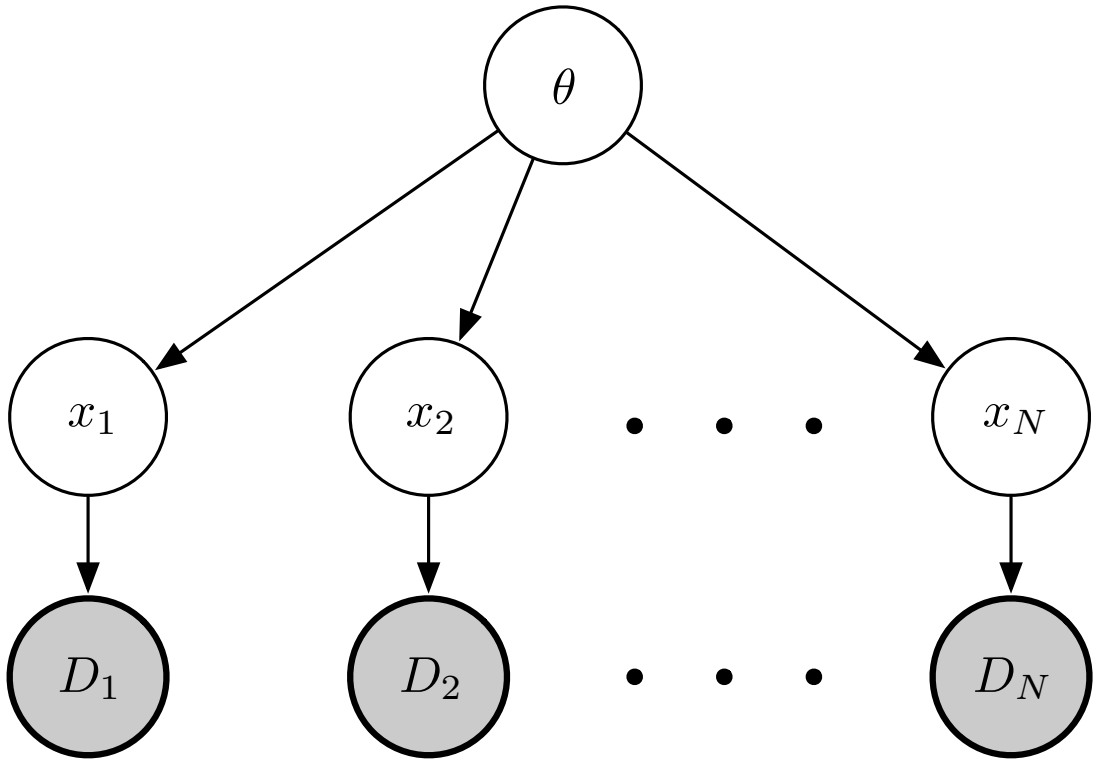
with $\hat{\theta}$ determined by the remaining data.

$f(x_1|\hat{\theta})$ behaves like a prior that shifts the x_1 estimate away from the peak of $\ell_1(x_i)$.

This generalizes the corrections derived by Eddington, Malmquist and Lutz-Kelker.

Landy & Szalay (1992) proposed adaptive Malmquist corrections that can be viewed as an approximation to this.

Graphical representation



$$\begin{aligned}\mathcal{L}(\theta, \{x_i\}) &\equiv p(\{D_i\}|\theta, \{x_i\}) \\ &= \prod_i p(D_i|x_i) f(x_i|\theta) = \prod_i \ell_i(x_i) f(x_i|\theta)\end{aligned}$$

A two-level *multi-level model* (MLM).

Bayesian MLMs in Astronomy

Surveys (number counts/"log N–log S"/Malmquist):

- GRB peak flux dist'n (Loredo & Wasserman 1998)
- TNO/KBO magnitude distribution (Gladman$^+$ 1998; Petit$^+$ 2008)
- MLM tutorial; Malmquist-type biases in cosmology (*Loredo & Hendry 2009* in *BMIC* book)
- "Extreme deconvolution" for proper motion surveys (Bovy, Hogg, & Roweis 2011)

Directional & spatio-temporal coincidences:

- GRB repetition (Luo$^+$ 1996; Graziani$^+$ 1996)
- GRB host ID (Band 1998; Graziani$^+$ 1999)
- VO cross-matching (Budavári & Szalay 2008)

Time series:

- SN 1987A neutrinos, uncertain energy vs. time (Loredo & Lamb 2002)
- Multivariate "Bayesian Blocks" (Dobigeon, Tourneret & Scargle 2007)
- SN Ia multicolor light curve modeling (Mandel^{+} 2009, 2011)

Linear regression with measurement error:

- QSO hardness vs. luminosity (Kelly 2007, 2011)

See 2010 CASt Surveys School notes for more!

`http://astrostatistics.psu.edu/su10/surveys.html`

Key Examples

1. Simple examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
2. Multilevel models for measurement error
3. **Bayesian calculation**
4. Bayesian Software
5. Closing Reflections

Statistical Integrals

Inference with independent data

Consider N data, $D = \{x_i\}$; and model M with m parameters.

Suppose $\mathcal{L}(\theta) = p(x_1|\theta)\, p(x_2|\theta) \cdots p(x_N|\theta)$.

Frequentist integrals

Find long-run properties of procedures via sample space integrals:

$$\mathcal{I}(\theta) = \int dx_1\, p(x_1|\theta) \int dx_2\, p(x_2|\theta) \cdots \int dx_N\, p(x_N|\theta) f(D, \theta)$$

Rigorous analysis must explore the θ dependence; rarely done in practice.

"Plug-in" approximation: Report properties of procedure for $\theta = \hat{\theta}$. *Asymptotically* accurate (for large N, expect $\hat{\theta} \to \theta$).

"Plug-in" results are easy via Monte Carlo (due to independence).

Bayesian integrals

$$\int d^m\theta\, g(\theta)\, p(\theta|M)\, \mathcal{L}(\theta) \quad = \quad \int d^m\theta\, g(\theta)\, \overbrace{q(\theta)} \, p(\theta|M)\, \mathcal{L}(\theta)$$

- $g(\theta) = 1 \to p(D|M)$ (norm. const., model likelihood)
- $g(\theta) =$ 'box' $\to$ credible region
- $g(\theta) = \theta \to$ posterior mean for θ

Such integrals are sometimes easy if analytic (especially in low dimensions), often easier than frequentist counterparts (e.g., normal credible regions, Student's t).

Asymptotic approximations: Require ingredients familiar from frequentist calculations. Bayesian calculation is *not significantly harder* than frequentist calculation in this limit.

Numerical calculation: For "large" m (> 4 is often enough!) the integrals are often very challenging because of structure (e.g., correlations) in parameter space. This is usually pursued *without making any procedural approximations*.

Bayesian Computation

Large sample size: Laplace approximation

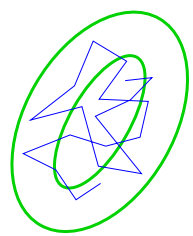
- Approximate posterior as multivariate normal $\rightarrow$ det(covar) factors
- Uses ingredients available in χ^2/ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$

Modest-dimensional models ($d \lesssim 10$ to 20)

- Adaptive cubature
- Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($d \gtrsim 5$)

- Posterior sampling — create RNG that samples posterior
- MCMC is most general framework — *Murali's lab*



See SCMA 5 Bayesian Computation tutorial notes for more!

Key Examples

1. Simple examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
2. Multilevel models for measurement error
3. Bayesian calculation
4. **Bayesian Software**
5. Closing Reflections

Tools for Computational Bayes

Astronomer/Physicist Tools

- **BIE** `http://www.astro.umass.edu/~weinberg/BIE/`
  Bayesian Inference Engine: General framework for Bayesian inference, tailored to astronomical and earth-science survey data. Built-in database capability to support analysis of terabyte-scale data sets. Inference is by Bayes via MCMC.
- **CIAO/Sherpa** `http://cxc.harvard.edu/sherpa/`
  On/off marginal likelihood support, and Bayesian Low-Count X-ray Spectral (BLoCXS) analysis via MCMC via the **pyblocxs** extension `https://github.com/brefsdal/pyblocxs`
- **XSpec** `http://heasarc.nasa.gov/xanadu/xspec/`
  Includes some basic MCMC capability
- **CosmoMC** `http://cosmologist.info/cosmomc/`
  Parameter estimation for cosmological models using CMB, etc., via MCMC
- **MultiNest** `http://ccpforge.cse.rl.ac.uk/gf/project/multinest/`
  Bayesian inference via an approximate implementation of the nested sampling algorithm
- **ExoFit** `http://www.homepages.ucl.ac.uk/~ucapola/exofit.html`
  Adaptive MCMC for fitting exoplanet RV data
- **extreme-deconvolution**
  `http://code.google.com/p/extreme-deconvolution/`
  Multivariate density estimation with measurement error, via a multivariate normal finite mixture model; partly Bayesian; Python & IDL wrappers

Astronomer/Physicist Tools, cont'd...

- **root/RooStats** `https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome`
  Statistical tools for particle physicists; Bayesian support being incorporated
- **CDF Bayesian Limit Software**
  `http://www-cdf.fnal.gov/physics/statistics/statistics_software.html`
  Limits for Poisson counting processes, with background & efficiency uncertainties
- **SuperBayeS** `http://www.superbayes.org/`
  Bayesian exploration of supersymmetric theories in particle physics using the MultiNest algorithm; includes a MATLAB GUI for plotting
- **CUBA** `http://www.feynarts.de/cuba/`
  Multidimensional integration via adaptive cubature, adaptive importance sampling & stratification, and QMC (C/C++, Fortran, and Mathematica; R interface also via 3rd-party R2Cuba)
- **Cubature** `http://ab-initio.mit.edu/wiki/index.php/Cubature`
  Subregion-adaptive cubature in C, with a 3rd-party R interface; intended for low dimensions (< 7)
- **APEMoST** `http://apemost.sourceforge.net/doc/`
  Automated Parameter Estimation and Model Selection Toolkit in C, a general-purpose MCMC environment that includes parallel computing support via MPI; motivated by asteroseismology problems
- **Inference** Forthcoming at `http://inference.astro.cornell.edu/`
  Several self-contained Bayesian modules; Parametric Inference Engine

Python

- **PyMC** `http://code.google.com/p/pymc/`
  A framework for MCMC via Metropolis-Hastings; also implements Kalman filters and Gaussian processes. Targets biometrics, but is general.
- **SimPy** `http://simpy.sourceforge.net/`
  SimPy (rhymes with "Blimpie") is a process-oriented public-domain package for discrete-event simulation.
- **RSPython** `http://www.omegahat.org/`
  Bi-directional communication between Python and R
- **MDP** `http://mdp-toolkit.sourceforge.net/`
  Modular toolkit for Data Processing: Current emphasis is on machine learning (PCA, ICA...). Modularity allows combination of algorithms and other data processing elements into "flows."
- **Orange** `http://www.ailab.si/orange/`
  Component-based data mining, with preprocessing, modeling, and exploration components. Python/GUI interfaces to $C++$ implementations. Some Bayesian components.
- **ELEFANT** `http://rubis.rsise.anu.edu.au/elefant`
  Machine learning library and platform providing Python interfaces to efficient, lower-level implementations. Some Bayesian components (Gaussian processes; Bayesian ICA/PCA).

R and S

- **CRAN Bayesian task view**
  `http://cran.r-project.org/web/views/Bayesian.html`
  Overview of many R packages implementing various Bayesian models and methods; pedagogical packages; packages linking R to other Bayesian software (BUGS, JAGS)
- **Omega-hat** `http://www.omegahat.org/`
  RPython, RMatlab, R-Xlisp
- **BOA** `http://www.public-health.uiowa.edu/boa/`
  Bayesian Output Analysis: Convergence diagnostics and statistical and graphical analysis of MCMC output; can read BUGS output files.
- **CODA**
  `http://www.mrc-bsu.cam.ac.uk/bugs/documentation/coda03/cdaman03.html`
  Convergence Diagnosis and Output Analysis: Menu-driven R/S plugins for analyzing BUGS output
- **R2Cuba**
  `http://w3.jouy.inra.fr/unites/miaj/public/logiciels/R2Cuba/welcome.html`
  R interface to Thomas Hahn's `Cuba` library (see above) for deterministic and Monte Carlo cubature

Java

- **Hydra** `http://research.warnes.net/projects/mcmc/hydra/`
  HYDRA provides methods for implementing MCMC samplers using Metropolis, Metropolis-Hastings, Gibbs methods. In addition, it provides classes implementing several unique adaptive and multiple chain/parallel MCMC methods.
- **YADAS** `http://www.stat.lanl.gov/yadas/home.html`
  Software system for statistical analysis using MCMC, based on the multi-parameter Metropolis-Hastings algorithm (rather than parameter-at-a-time Gibbs sampling)
- **Omega-hat** `http://www.omegahat.org/`
  Java environment for statistical computing, being developed by XLisp-stat and R developers

C/C++/Fortran

- **BayeSys 3** `http://www.inference.phy.cam.ac.uk/bayesys/`
  Sophisticated suite of MCMC samplers including transdimensional capability, by the author of MemSys
- **fbm** `http://www.cs.utoronto.ca/~radford/fbm.software.html`
  Flexible Bayesian Modeling: MCMC for simple Bayes, nonparametric Bayesian regression and classification models based on neural networks and Gaussian processes, and Bayesian density estimation and clustering using mixture models and Dirichlet diffusion trees
- **BayesPack, DCUHRE**
  `http://www.sci.wsu.edu/math/faculty/genz/homepage`
  Adaptive quadrature, randomized quadrature, Monte Carlo integration
- **BIE**, **CDF Bayesian limits**, **CUBA** (see above)

Other Statisticians' & Engineers' Tools

- **BUGS/WinBUGS** `http://www.mrc-bsu.cam.ac.uk/bugs/`
  Bayesian Inference Using Gibbs Sampling: Flexible software for the Bayesian analysis of complex statistical models using MCMC
- **OpenBUGS** `http://mathstat.helsinki.fi/openbugs/`
  BUGS on Windows and Linux, and from inside the R
- **JAGS** `http://www-fis.iarc.fr/~martyn/software/jags/`
  "Just Another Gibbs Sampler;" MCMC for Bayesian hierarchical models
- **XLisp-stat** `http://www.stat.uiowa.edu/~luke/xls/xlsinfo/xlsinfo.html`
  Lisp-based data analysis environment, with an emphasis on providing a framework for exploring the use of dynamic graphical methods
- **ReBEL** `http://choosh.csee.ogi.edu/rebel/`
  Library supporting recursive Bayesian estimation in Matlab (Kalman filter, particle filters, sequential Monte Carlo).

Key Examples

1. Simple examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
2. Multilevel models for measurement error
3. Bayesian calculation
4. Bayesian Software
5. **Closing Reflections**

Closing Reflections

Philip Dawid (2000)

> What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle. There is no room for clever tricks or an alphabetic cornucopia of definitions and optimality criteria. I have heard people use this 'dullness' as an argument against Bayesianism. One might as well complain that Newton's dynamics, being based on three simple laws of motion and one of gravitation, is a poor substitute for the richness of Ptolemy's epicyclic system.
>
> All my experience teaches me that it is invariably more fruitful, and leads to deeper insights and better data analyses, to explore the consequences of being a 'thoroughly boring Bayesian'.

Dennis Lindley (2000)

> The philosophy places more emphasis on model construction than on formal inference. . . I do agree with Dawid that 'Bayesian statistics is fundamentally boring'. . . My only qualification would be that the theory may be boring but the applications are exciting.