Introduction to Bayesian Inference

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

June 10, 2006

Outline

Outline

Scientific Method

Science is more than a body of knowledge; it is a way of thinking. The method of science, as stodgy and grumpy as it may seem, is far more important than the findings of science.

—Carl Sagan

Scientists *argue!*

Argument $\equiv$ Collection of statements comprising an act of reasoning from premises to a conclusion

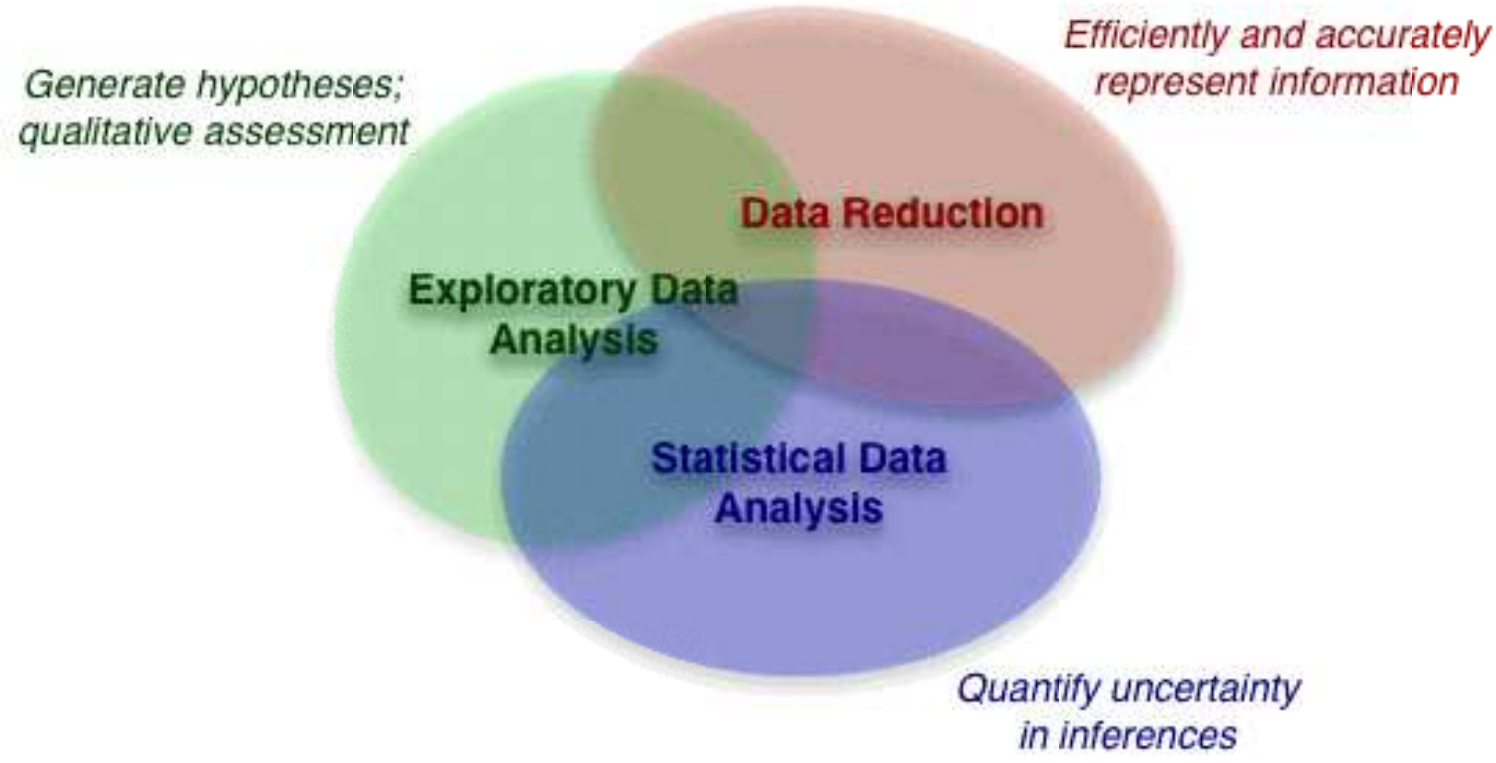
A key goal of science: Explain or predict *quantitative measurements*

Framework: Mathematical modeling

$\Rightarrow$ Science uses rational argument to construct and appraise mathematical models for measurements

Data Analysis

Building & Appraising Arguments Using Data



Statistical inference is but one of several interacting modes of analyzing data.

Statistics as Principled Argument

[My aim is] to locate the field of statistics with respect to rhetoric and narrative. My central theme is that good statistics involves principled argument that conveys an interesting and credible point.

Data analysis should not be pointlessly formal. It should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations.

I have proposed that the proper function of statistics is to formulate good arguments explaining comparative differences, hopefully in an interesting way.

—Robert Abelson, *Statistics as Principled Argument*

Bayesian Statistical Inference

- A different approach to *all* statistical inference problems (i.e., not just another method in the list: BLUE, maximum likelihood, χ^2 testing, ANOVA, survival analysis . . .)
- Foundation: Use probability theory to quantify the strength of arguments (i.e., a more abstract view than restricting PT to describe variability in repeated "random" experiments)
- Focuses on *deriving consequences of modeling assumptions* rather than *devising and calibrating procedures*

Hypotheses, Data, and Models

We seek to appraise scientific hypotheses in light of observed data and modeling assumptions.

Consider the data and modeling assumptions to be the premises of an argument with each of various hypotheses, H_i, as conclusions: $H_i|D_{\text{obs}}, I$. ($I =$ "background information," everything deemed relevant besides the observed data)

$P(H_i|D_{\text{obs}}, I)$ measures the degree to which (D_{obs}, I) support H_i. It provides an ordering among the H_i (and among other hypotheses derivable from them).

Probability theory tells us how to analyze and appraise the argument, i.e., how to calculate $P(H_i|D_{\text{obs}}, I)$ from simpler, hopefully more accessible probabilities.

Outline

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information using the rules of probability theory.

Probability Theory Axioms:

'OR' (sum rule) $\quad P(H_1 + H_2|I) = P(H_1|I) + P(H_2|I) - P(H_1, H_2|I)$

'AND' (product rule) $\quad P(H_1, D|I) = P(H_1|I)\, P(D|H_1, I)$
$= P(D|I)\, P(H_1|D, I)$

Three Important Theorems

Bayes's Theorem (BT)

Consider $P(H_i, D_{\text{obs}}|I)$ using the product rule:

$$\begin{aligned} P(H_i, D_{\text{obs}}|I) &= P(H_i|I)\, P(D_{\text{obs}}|H_i, I) \\ &= P(D_{\text{obs}}|I)\, P(H_i|D_{\text{obs}}, I) \end{aligned}$$

Solve for the *posterior probability*:

$$P(H_i|D_{\text{obs}}, I) = P(H_i|I)\, \frac{P(D_{\text{obs}}|H_i, I)}{P(D_{\text{obs}}|I)}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$$

norm. const. $P(D_{\text{obs}}|I)$ = prior predictive

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ (I asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i|I) &= \sum_i P(B_i|A, I)P(A|I) = P(A|I) \\ &= \sum_i P(B_i|I)P(A|B_i, I)\end{aligned}$$

If we do not see how to get $P(A|I)$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|I) = \sum_i P(B_i|I)P(A|B_i, I)$$

If our problem already has B_i in it, we can use LTP to get $P(A|I)$ from the joint probabilities—*marginalization*:

$$P(A|I) = \sum_i P(A, B_i|I)$$

Example: Take $A = D_{\rm obs}$, $B_i = H_i$; then

$$\begin{aligned} P(D_{\rm obs}|I) &= \sum_i P(D_{\rm obs}, H_i|I) \\ &= \sum_i P(H_i|I) P(D_{\rm obs}|H_i, I) \end{aligned}$$

prior predictive for $D_{\rm obs}$ = Average likelihood for H_i
(*aka* "marginal likelihood")

Normalization

For *exclusive, exhaustive* H_i,

$$\sum_i P(H_i|\cdots) = 1$$

Recap

Bayesian inference is more than BT

Bayesian inference quantifies uncertainty by reporting probabilities for things we are uncertain of, given specified premises.

It uses *all* of probability theory, not just (or even primarily) Bayes's theorem.

The Rules in Plain English

- Ground rule: Specify premises that include everything relevant that you know or are willing to presume to be true (for the sake of the argument!).
- BT: Make your appraisal account for all of your premises.

  Things you know are false must not enter your accounting.
- LTP: If the premises allow multiple arguments for a hypothesis, its appraisal must account for all of them.

  Do not just focus on the most or least favorable way a hypothesis may be realized.

Outline

Inference With Parametric Models

Models M_i ($i = 1$ to N), each with parameters θ_i, each imply a *sampling dist'n* (conditional predictive dist'n for possible data):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the observed data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty).

Three Classes of Problems

Parameter Estimation

Premise = choice of model (pick specific i)
$\longrightarrow$ What can we say about θ_i?

Model Assessment

- Model comparison: Premise = $\{M_i\}$
  $\longrightarrow$ What can we say about i?
- Model adequacy/GoF: Premise = M_1
  $\longrightarrow$ Is M_1 adequate?

Model Averaging

Models share some common params: $\theta_i = \{\phi, \eta_i\}$
$\longrightarrow$ What can we say about ϕ?
(Systematic error is an example)

Parameter Estimation

Problem statement

I = Model M with parameters θ (+ any add'l info)
H_i = statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"
Probability for any such statement can be found using a *probability density function* (PDF) for θ:

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \cdots) &= f(\theta) d\theta \\ &= p(\theta | \cdots) d\theta \end{aligned}$$

Posterior probability density

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

Summaries of posterior

- "Best fit" values:
  - Mode, $\hat{\theta}$, maximizes $p(\theta|D, M)$
  - Posterior mean, $\langle\theta\rangle = \int d\theta\, \theta\, p(\theta|D, M)$
- Uncertainties:
  - Credible region Δ of probability C:
    $C = P(\theta \in \Delta|D, M) = \int_\Delta d\theta\, p(\theta|D, M)$
    Highest Posterior Density (HPD) region has $p(\theta|D, M)$ higher inside than outside
  - Posterior standard deviation, variance, covariances
- Marginal distributions
  - Interesting parameters ψ, nuisance parameters ϕ
  - Marginal dist'n for ψ: $p(\psi|D, M) = \int d\phi\, p(\psi, \phi|D, M)$

Nuisance Parameters and Marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example: The data measure a rate r that is a sum of an interesting signal s and a background b. We have additional data just about b.

What do the data tell us about s?

Marginal posterior distribution

$$\begin{aligned} p(s|D,M) &= \int db\; p(s,b|D,M) \\ &\propto p(s|M) \int db\; p(b|s)\, \mathcal{L}(s,b) \\ &\equiv p(s|M)\mathcal{L}_m(s) \end{aligned}$$

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

$$\mathcal{L}_m(s) \approx \mathcal{L}[s, \hat{b}(s)]\, \delta b(s)$$

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

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}, \quad \sigma_s^2 = \sigma_r^2 + \sigma_b^2$
Background *subtraction* is a special case of background *marginalization*.

Model Comparison

Problem statement

$I = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models.
$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model

$$\begin{aligned} p(M_i|D,I) &= p(M_i|I)\frac{p(D|M_i,I)}{p(D|I)} \\ &\propto p(M_i|I)\mathcal{L}(M_i) \end{aligned}$$

But $\mathcal{L}(M_i) = p(D|M_i) = \int d\theta_i \, p(\theta_i|M_i)p(D|\theta_i, M_i)$.

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Varied terminology: Prior predictive = Average likelihood = Global likelihood = Marginal likelihood = (Weight of) Evidence for model

Odds and Bayes factors

Ratios of probabilities for two propositions using the same premises are called *odds*:

$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D,I)}{p(M_j|D,I)} \\ &= \frac{p(M_i|I)}{p(M_j|I)} \times \frac{p(D|M_j,I)}{p(D|M_j,I)} \end{aligned}$$

The data-dependent part is called the *Bayes factor*:

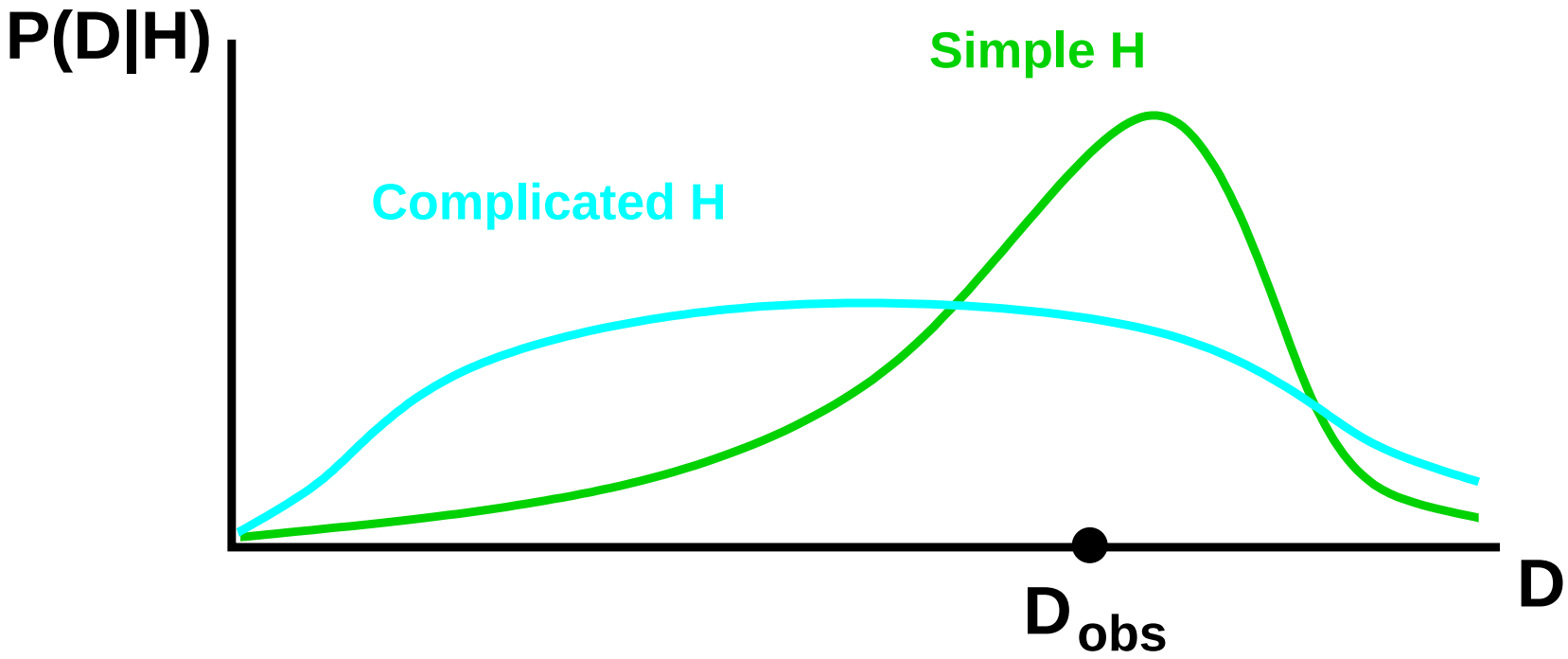
$$B_{ij} \equiv \frac{p(D|M_j,I)}{p(D|M_j,I)}$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods.

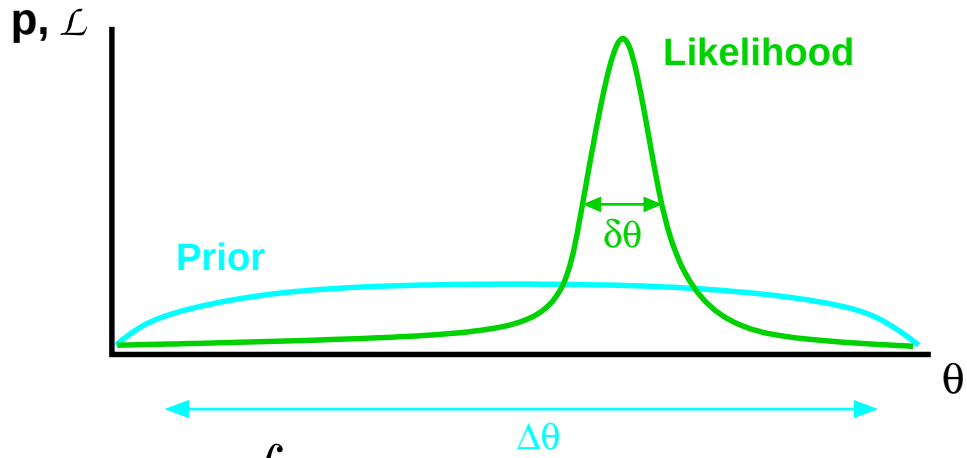
An Automatic Occam's Razor

Predictive probabilities can favor simpler models

$$p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i)$$



The Occam Factor



$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M)\; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Occam Factor}
\end{aligned}
$$

Models with more parameters often make the data more probable— *for the best fit*.
Occam factor penalizes models for "wasted" volume of parameter space.

Theme: Parameter Space Volume

Bayesian calculations sum/integrate over parameter/hypothesis space! (Frequentist calculations average over sample space.)

- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters.
- Model likelihoods have Occam factors resulting from parameter space volume factors.

Many virtues of Bayesian methods can be attributed to this accounting for the "size" of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added "by hand").

Outline

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 $\epsilon_i \sim N(0, \sigma^2)$.

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

Find dependence of Q on μ by completing the square:

$$\begin{aligned} Q &= \sum_i (d_i - \mu)^2 \\ &= \sum_i d_i^2 + N\mu^2 - 2N\mu\overline{d} \qquad \text{where } \overline{d} \equiv \frac{1}{N}\sum_i d_i \\ &= N(\mu - \overline{d})^2 + Nr^2 \quad \text{where } r^2 \equiv \frac{1}{N}\sum_i (d_i - \overline{d})^2 \end{aligned}$$

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 $w = \sigma/\sqrt{N}$.

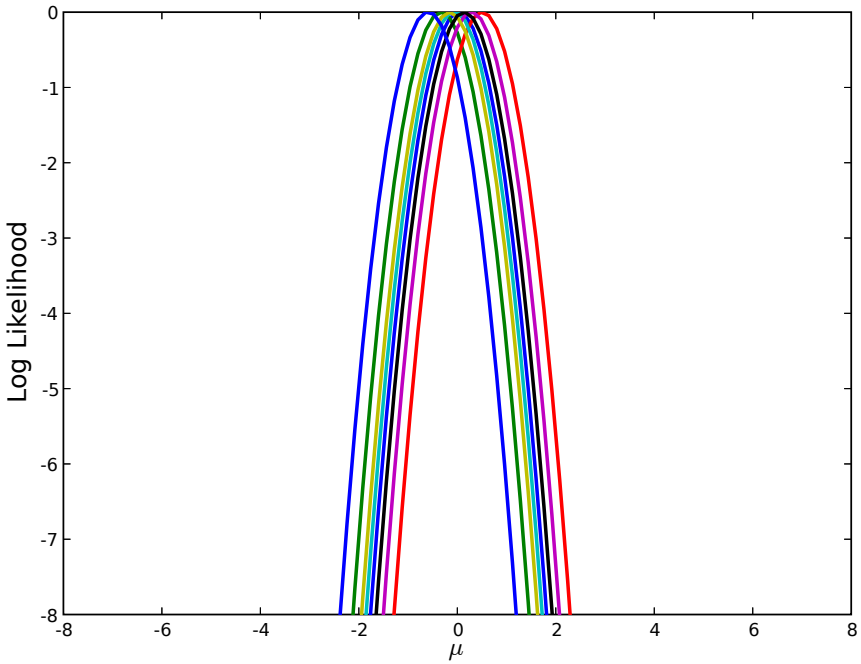
68.3% HPD credible region for μ is $\overline{d} \pm \sigma/\sqrt{N}$.

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

How Is This Different?

Coverage of a frequentist *confidence region* is found by integrating over *other samples*.

Likelihoods for 8 samples of size $N = 5$; for $\mu = 0$



Frequentist coverage of interval $I(D)$:

$$C(\mu) = \int dD\ [\mu \in I(D)]\, p(D|\mu)$$

Confidence level $CL = \min_\mu C(\mu)$

Probability in Bayesian credible region $I(D)$:

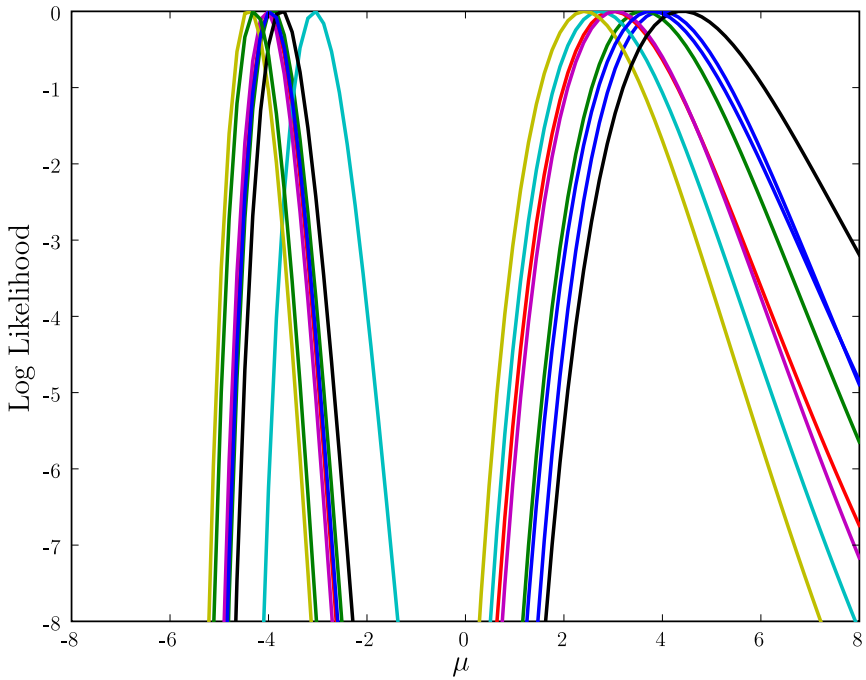
$$P = \frac{1}{Z} \int_{I(D)} d\mu\ p(\mu) pd(D|\mu)$$

Can show $P = \int d\mu\ p(\mu) C(\mu) \rightarrow$ Bayes intervals have exact *avg.* coverage.

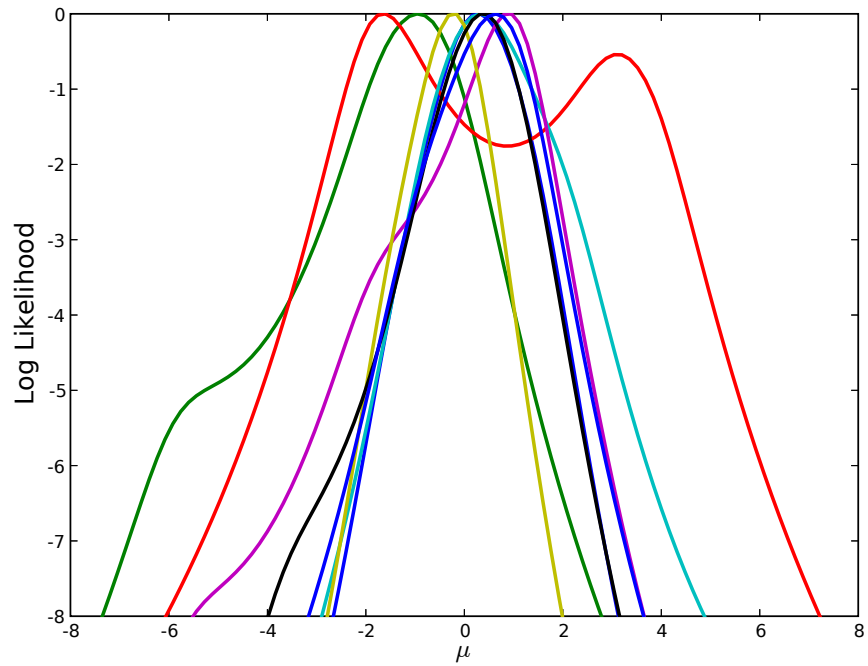
When Does This Matter?

- Varying σ, nonlinear models — Coverage depends on true value
- Other distributions — No sufficient statistics

Varying σ



Cauchy



Informative Conjugate Prior

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

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

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

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

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 $p(\mu|D, \sigma, M)$

Likelihood

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

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

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

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

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:

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

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 standard choices:

- 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:

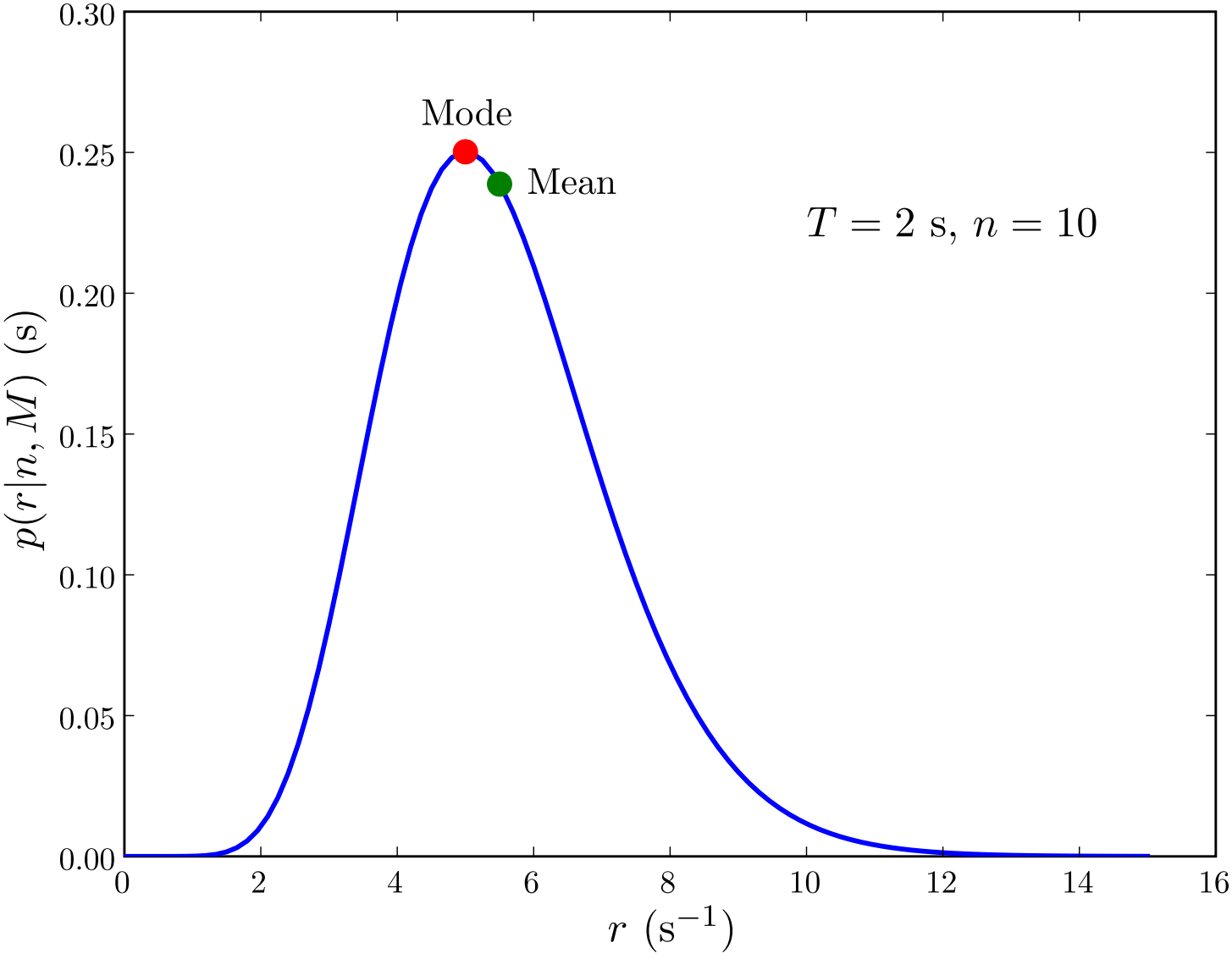
$$p_\Gamma(x|\alpha, s) = \frac{1}{s\Gamma(\alpha)} \left(\frac{x}{s}\right)^{\alpha-1} e^{-x/s}$$

Moments:

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

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 $r \rightarrow p(r|I) = C$
Require (discrete) predictive distribution to be flat:

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

A convention

- 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

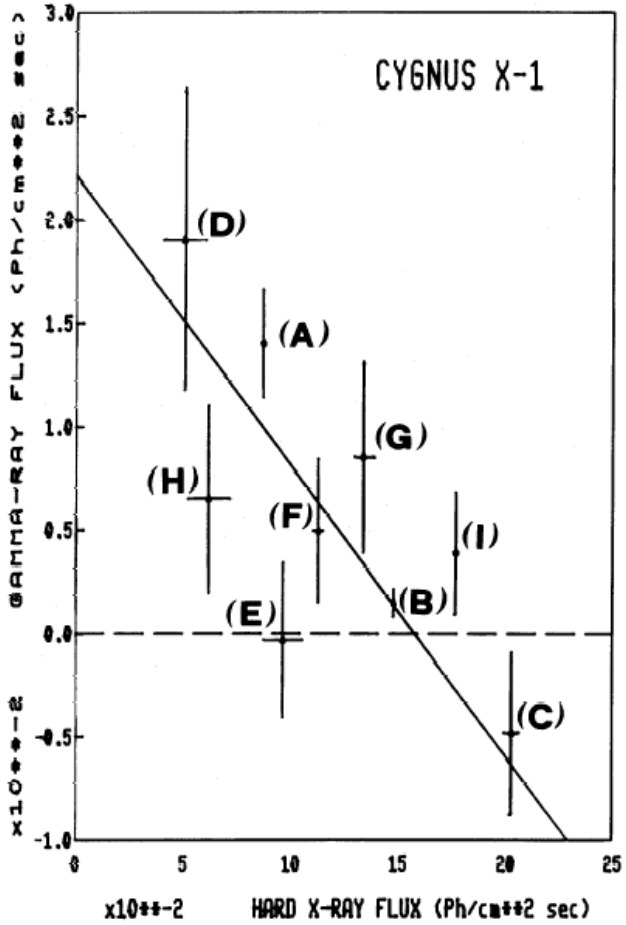
$$\begin{aligned} \hat{b} &= N_{\rm off}/T_{\rm off}; & \sigma_b &= \sqrt{N_{\rm off}}/T_{\rm off} \\ \hat{r} &= N_{\rm on}/T_{\rm on}; & \sigma_r &= \sqrt{N_{\rm on}}/T_{\rm on} \\ \hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2} \end{aligned}$$

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

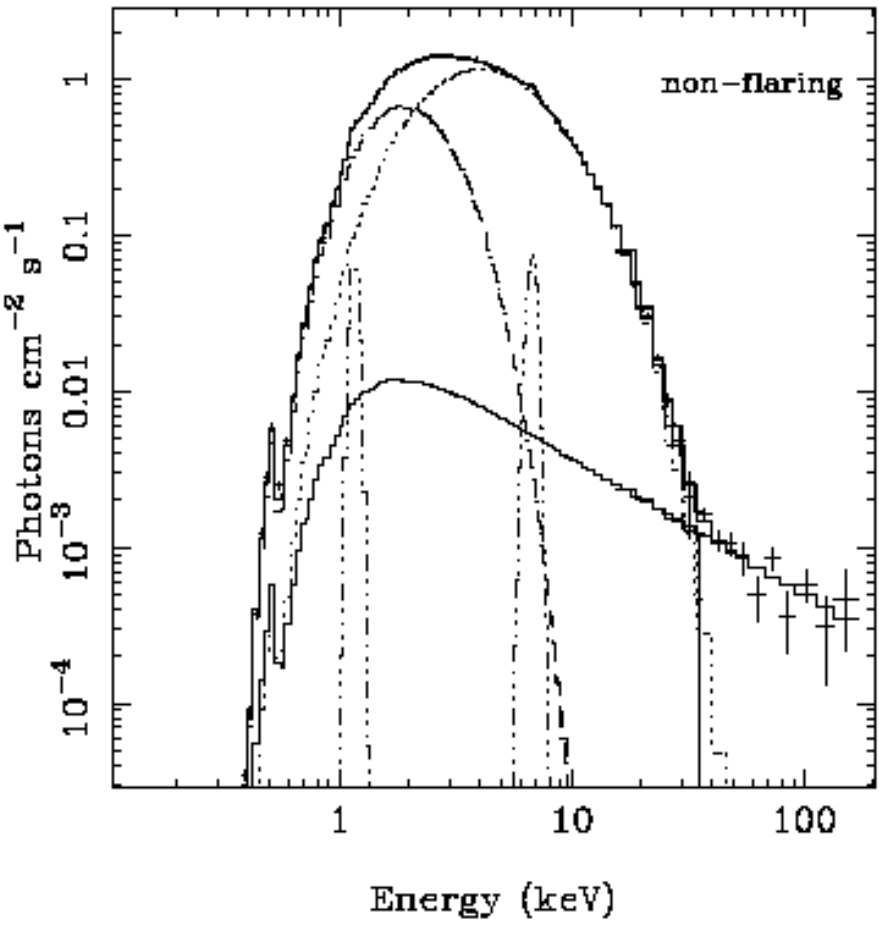
Examples

Spectra of X-Ray Sources

Bassani et al. 1989

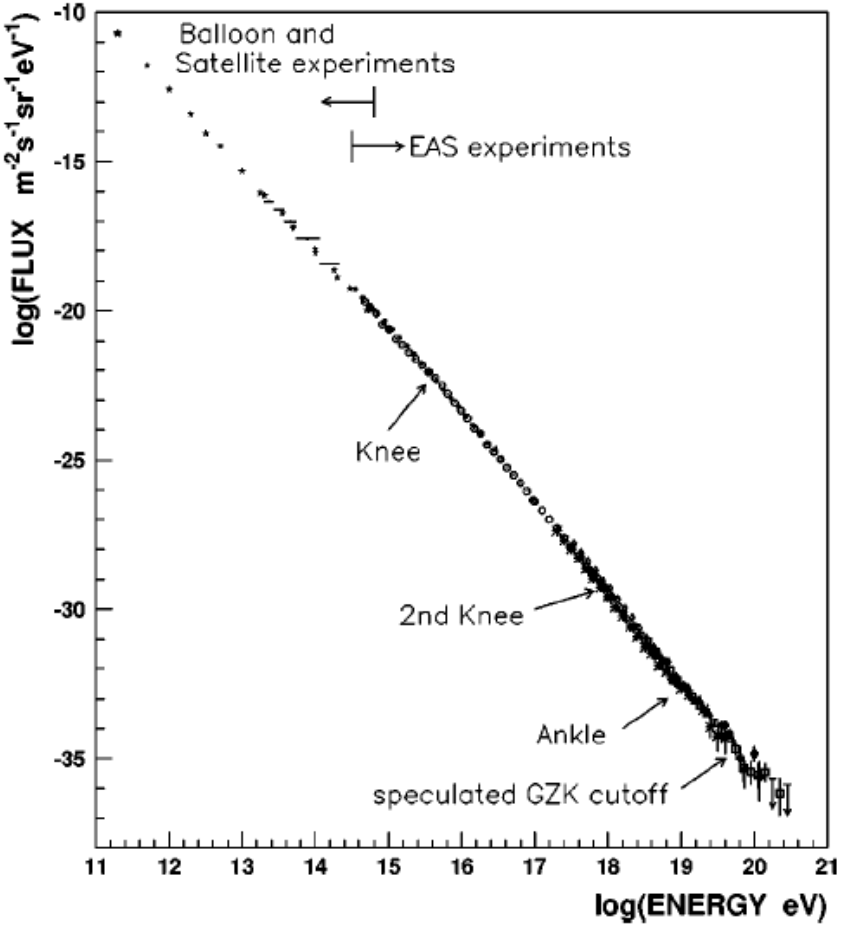


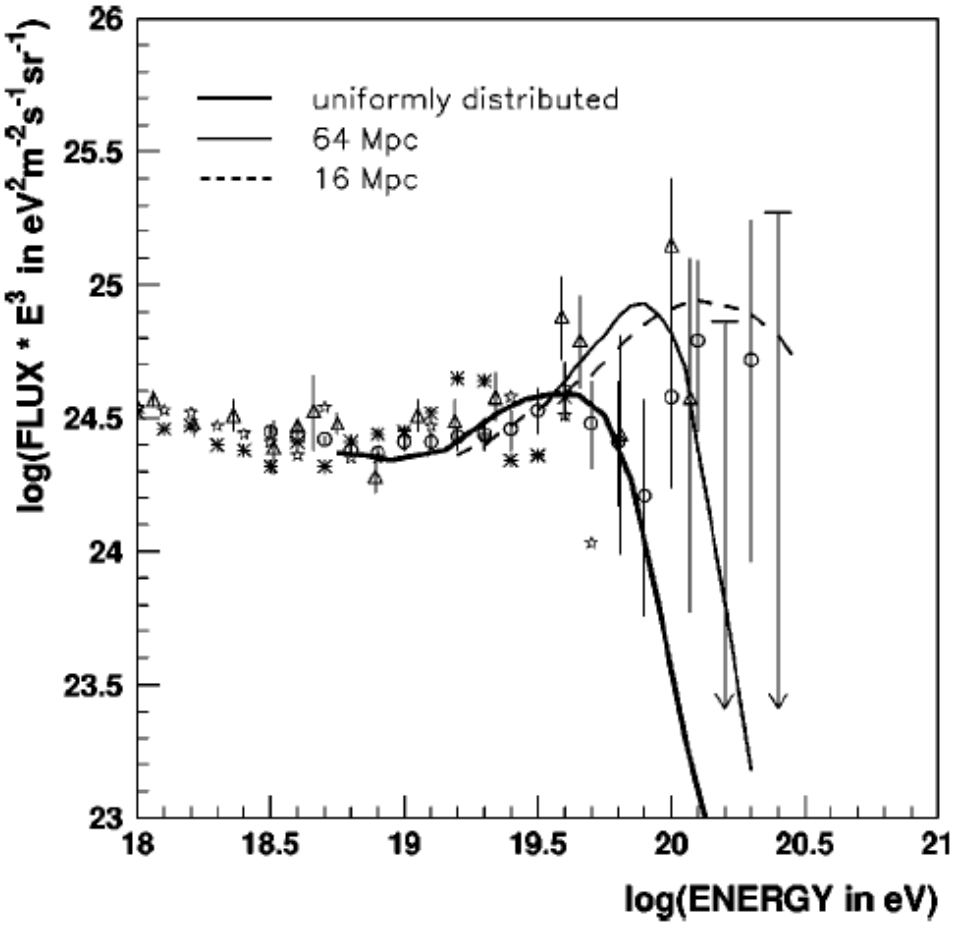
Di Salvo et al. 2001



Spectrum of Ultrahigh-Energy Cosmic Rays

Nagano & Watson 2000





Backgrounds as Nuisance Parameters

Background marginalization with Gaussian noise

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):

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

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

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;

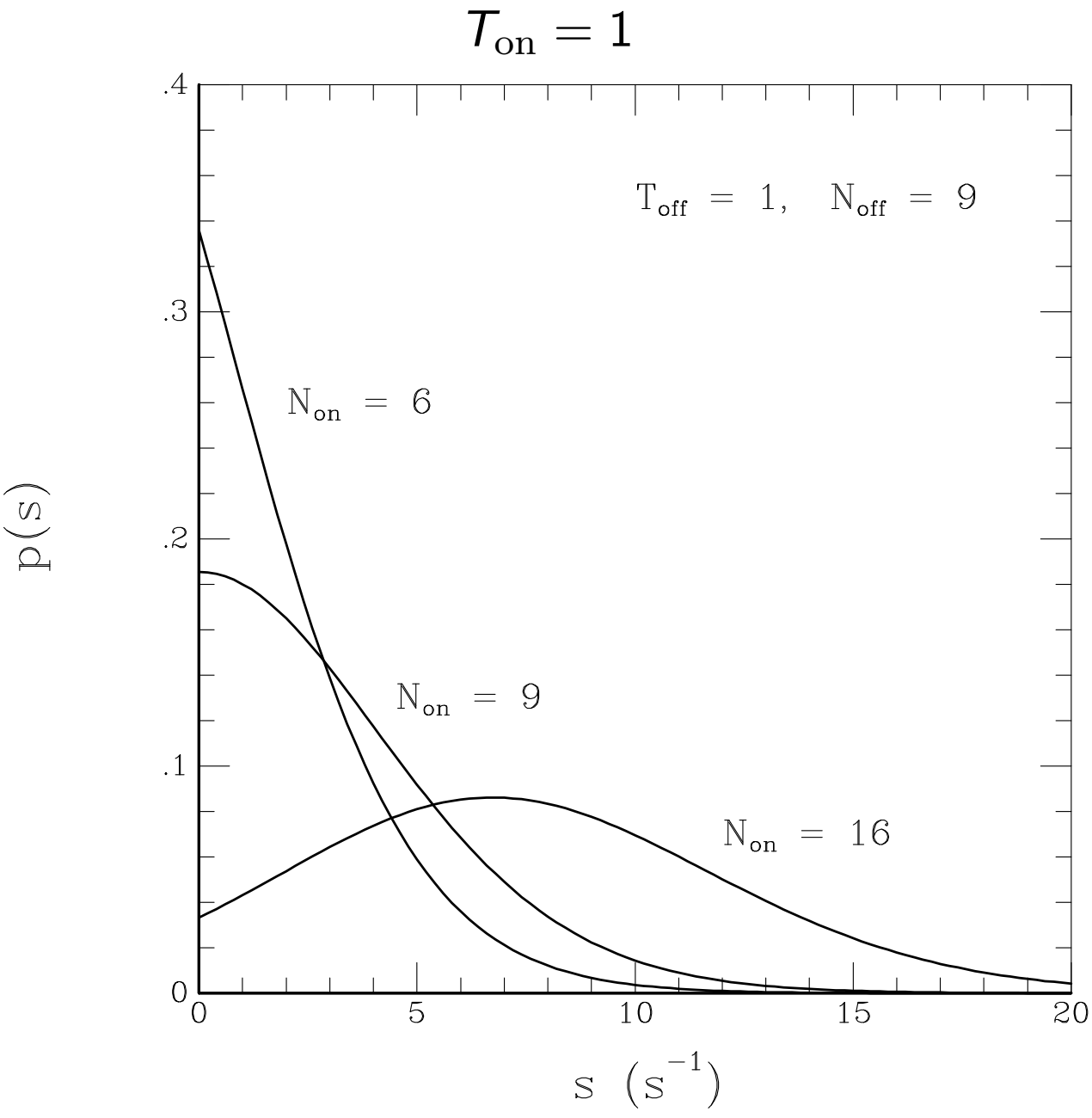
$$\begin{aligned} p(s|N_{\rm on}, I_{\rm all}) &= \int db\; p(s, b \mid N_{\rm on}, I_{\rm all}) \\ &\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})} \end{aligned}$$

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

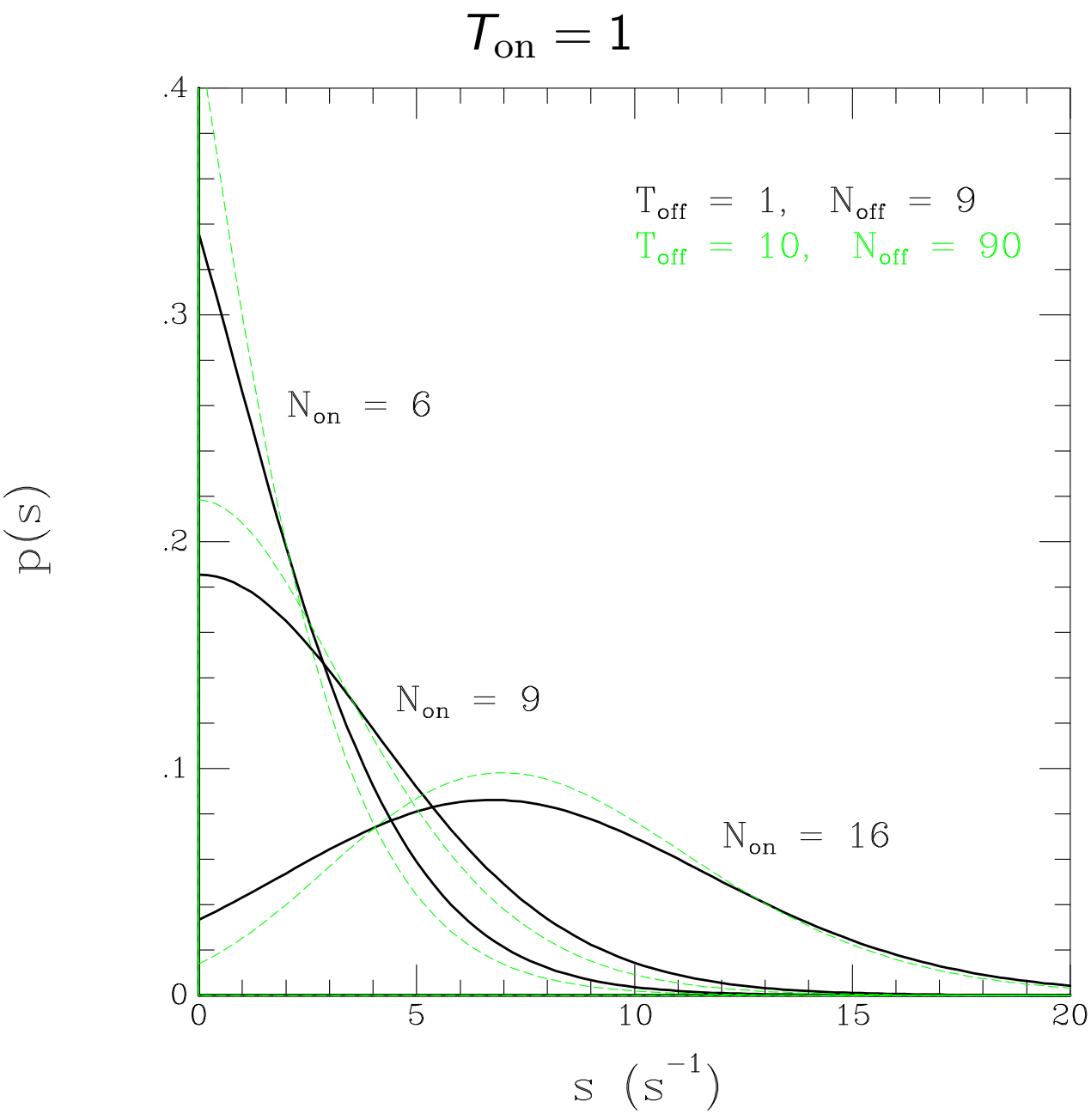
$$\begin{aligned} p(s|N_{\rm on}, I_{\rm all}) &= \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!} \\ C_i &\propto \left(1 + \frac{T_{\rm off}}{T_{\rm on}}\right)^i \frac{(N_{\rm on}+N_{\rm off}-i)!}{(N_{\rm on}-i)!} \end{aligned}$$

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



Multibin On/Off

The more typical on/off scenario:

Data = spectrum or image with counts in many bins

Model M gives signal rate $s_k(\theta)$ in bin k, parameters θ

To infer θ, we need the likelihood:

$$\mathcal{L}(\theta) = \prod_k p(N_{\mathrm{on}\,k}, N_{\mathrm{off}\,k} | s_k(\theta), M)$$

For each k, we have an on/off problem as before, only we just need the marginal likelihood for s_k (not the posterior). The same C_i coefficients arise.

XSPEC and CIAO/Sherpa provide this as an option.

CHASC approach does the same thing via data augmentation.

Bayesian Computation

Large N—Laplace approximation

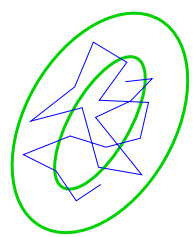
- Approximate posterior as (multivariate) normal
- Uses ingredients available in χ^2/ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$

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

- Adaptive quadrature
- Monte Carlo integration (importance sampling, QMC)

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

- Posterior sampling—create RNG that samples posterior
- MCMC is most general framework



Outline

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:

- General relationships between probability and frequency
- Long-run performance of Bayesian procedures

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

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

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:

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

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

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

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.

Infer probability from related trials

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

A formalism that distinguishes p from f from the outset is particularly valuable for exploring subtle connections. E.g., shrinkage is explored via hierarchical and empirical Bayes.

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.
- 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.

(Not so clear in nonparametric problems.)

Conclusion: Key Ideas

- Probability as generalized logic for appraising arguments
- Three theorems: BT, LTP, Normalization
- Calculations characterized by parameter space integrals
  - Credible regions, posterior expectations
  - Marginalization over nuisance parameters
  - Occam's razor via marginal likelihoods
- Probability & frequency
  - $p \leftrightarrow f$ in i.i.d. setting; otherwise more subtle
  - Parametric Bayesian procedures have good frequentist performance