Introduction to Bayesian inference: Key examples

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

IAC Winter School, 3–4 Nov 2014

Key examples: 3 sampling distributions

❶ **Binomial distribution (probability & frequency)**

❷ **Normal distribution (additive noise)**

❸ **Poisson distribution (rates & counts)**

Supplement

- Binary classification with binary data
- Negative binomial distribution, stopping rules
- Likelihood principle
- Relationships between probability & frequency

Key examples: 3 sampling distributions

1. **Binomial distribution (probability & frequency)**
2. Normal distribution (additive noise)
3. Poisson distribution (rates & counts)

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 (Bernoulli process):

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

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
- *Prior predictive ignorance:* Bayes's suggested "ignorance" here can mean that before doing the N trials, we have no preference for how many will be successes:

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

Consider the uniform prior 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

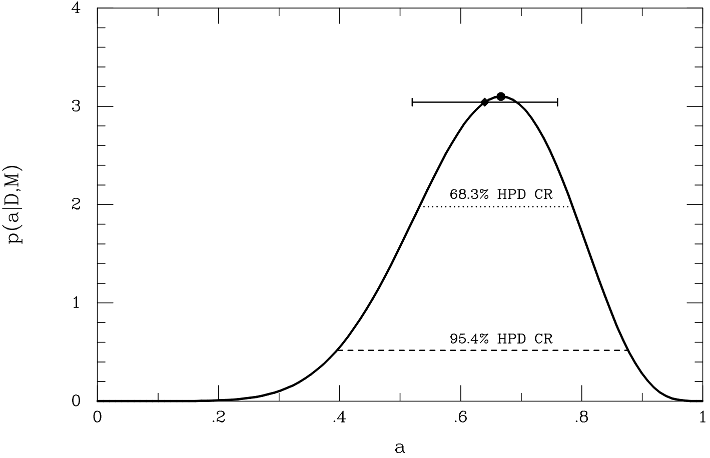
$$p(\alpha|D, M) \quad = \quad \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

A *Beta distribution*. Summaries:

- Best-fit: mode $\hat{\alpha} = \frac{n}{N} = 2/3$; $\langle\alpha\rangle = \frac{n+1}{N+2} \approx 0.64$
- 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*


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

Beta distribution (in general)

A two-parameter family of distributions for a quantity α in the unit interval $[0, 1]$:

$$p(\alpha|a, b) = \frac{1}{B(a, b)}\alpha^{a-1}(1-\alpha)^{b-1}$$

Summaries:

- Mode: $\hat{\alpha} = \frac{a-1}{(a-1)+(b-1)}$
- Mean: $\mu \equiv \mathrm{E}(\alpha) \equiv \langle\alpha\rangle = \frac{a}{a+b}$
- Variance: $\sigma^2 \equiv \mathrm{Var}(\alpha) = \frac{ab}{(a+b)^2(a+b+1)}$
- Cumulative distribution via incomplete beta function

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 (on the basis of best-fit α)

Bayes Factor (ratio of model likelihoods)

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

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

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.

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

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

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

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

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

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
(An example of the *likelihood principle*—see supplement)

The beta-binomial conjugate model

Generalize from the flat prior to a $\mathrm{Beta}(\alpha|a, b)$ prior for α

$$
\begin{aligned}
p(\alpha|n, M') &\propto \mathrm{Beta}(\alpha|a, b) \times \mathrm{Binom}(n|\alpha, N) \\
&\propto \alpha^{a-1}(1-\alpha)^{b-1} \times \alpha^{n}(1-\alpha)^{N-n} \\
&\propto \alpha^{n+a-1}(1-\alpha)^{N-n+b-1}
\end{aligned}
$$

$\Rightarrow$ the posterior is $\mathrm{Beta}(\alpha|n+a, N-n+b)$

When the prior and likelihood are such that the posterior is in the same family as the prior, the prior and likelihood are a *conjugate* pair

A Beta prior is a conjugate prior for both the binomial and Bernoulli process sampling distributions

Conjugacy $\rightarrow$ it's easy to chain inferences from multiple experiments

Key examples: 3 sampling distributions

1. Binomial distribution (probability & frequency)

2. **Normal distribution (additive noise)**

3. Poisson distribution (rates & counts)

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 with additive noise:

$$d_i = \mu + \epsilon_i, \qquad i = 1 \text{ to } N$$

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:

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

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
- *Reference prior:* Asymptotic information theory criterion $\Rightarrow p(\mu) \propto C$

This prior is *improper* unless bounded; formally we should bound it and take ∞ limit

(Minimal sample size arguments suggest impropriety is a *desirable* feature of uninformative priors)

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 turns out also to be normal

Posterior

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

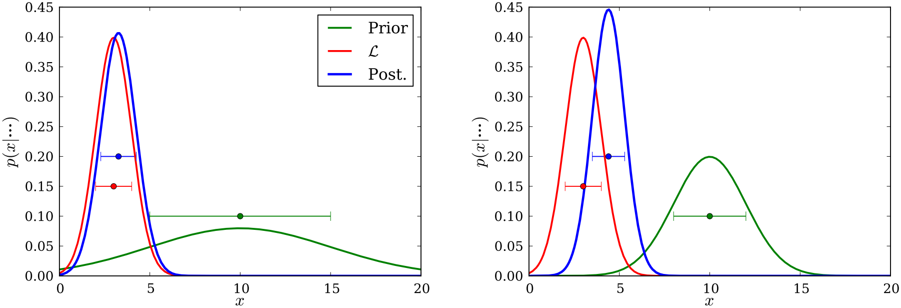
Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large; then

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

Principle of stable estimation/precise measurement — The prior affects estimates only when data are not informative relative to prior (J. Savage)

Conjugate normal examples:

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



Note we always have $\widetilde{w} < w$ (in the normal-normal setup)

Estimating a Normal Mean: Unknown σ

Supplement: Handling σ uncertainty by marginalizing over $\sigma \rightarrow$ *Student's t distribution* (heavier tails than normal)

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—but it won't always do that!

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/the Delta method—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 to 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 $g_\alpha(x; \psi)$:

$$d_i = \sum_{\alpha=1}^{M} A_\alpha g_\alpha(x_i; \psi) + \epsilon_i$$

$$\text{or} \qquad \vec{d} = \sum_{\alpha} A_\alpha \vec{g}_\alpha(\psi) + \vec{\epsilon}.$$

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; for sinusoidal g_α, generalizes periodograms

Key examples: 3 sampling distributions

1. Binomial distribution (probability & frequency)

2. Normal distribution (additive noise)

3. **Poisson distribution (rates & counts)**

Poisson Dist'n: Infer a Rate from Counts

Problem:

Observe n counts in T; infer rate, r

Likelihood

Poisson distribution:

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

See Jaynes, "Probability theory as logic" (MaxEnt 1990) for an instructive derivation

Prior

Two simple "uninformative" standard choices:

- r known to be *nonzero*: it is a scale parameter; scale invariance $\rightarrow$

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

  This corresponds to a flat prior on $\lambda = \log r$

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

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

The reference prior is $p(r|M) \propto 1/r^{1/2}$

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 λ (or inverse scale ϵ):

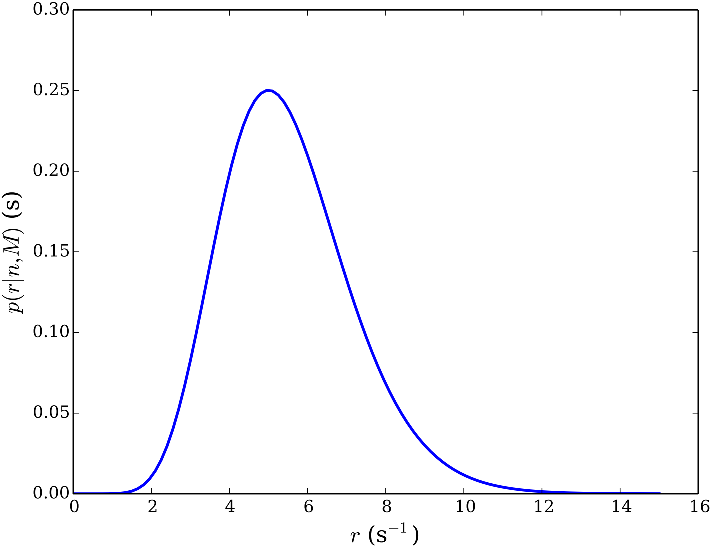
$$p_\Gamma(x|\alpha,\lambda) \equiv \frac{1}{\lambda\Gamma(\alpha)}\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-x/\lambda}$$
$$\equiv \frac{\epsilon}{\Gamma(\alpha)}(x\epsilon)^{\alpha-1}e^{-x\epsilon}$$

Moments:

$$\mathrm{E}(x) = \alpha\lambda = \frac{\alpha}{\epsilon} \qquad \mathrm{Var}(x) = \lambda^2\alpha = \frac{\alpha}{\epsilon^2}$$

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

- 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)



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

Conjugate prior

Note that a gamma distribution prior is the conjugate prior for the Poisson sampling distribution:

$$\begin{aligned} p(r|n, M') &\propto \mathrm{Gamma}(r|\alpha, \epsilon) \times \mathrm{Pois}(n|rT) \\ &\propto r^{\alpha-1} e^{-r\epsilon} \times r^n e^{-rT} \\ &\propto r^{\alpha+n-1} \exp[-r(T+\epsilon)] \end{aligned}$$

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 (use $\log r$ abscissa for scale parameter case)

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

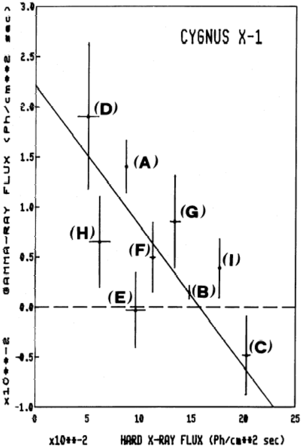
$$\begin{array}{ll} \hat{b} = N_{\rm off}/T_{\rm off}; & \sigma_b = \sqrt{N_{\rm off}}/T_{\rm off} \\ \hat{r} = N_{\rm on}/T_{\rm on}; & \sigma_r = \sqrt{N_{\rm on}}/T_{\rm on} \\ \hat{s} = \hat{r} - \hat{b}; & \sigma_s = \sqrt{\sigma_r^2 + \sigma_b^2} \end{array}$$

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

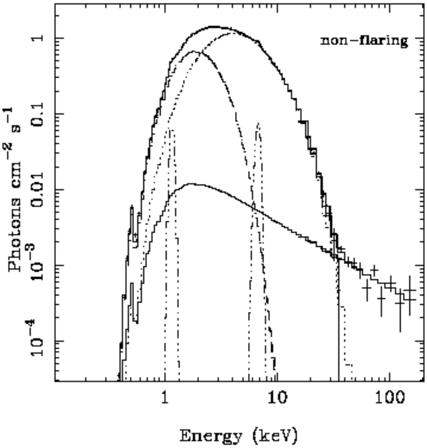
Examples

Spectra of X-Ray Sources

Bassani et al. 1989

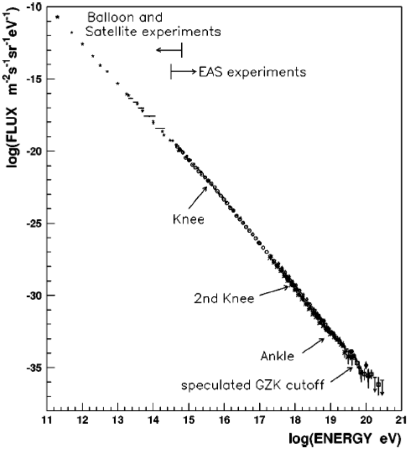


Di Salvo et al. 2001

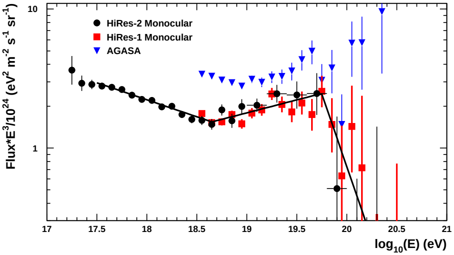


Spectrum of Ultrahigh-Energy Cosmic Rays

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;

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

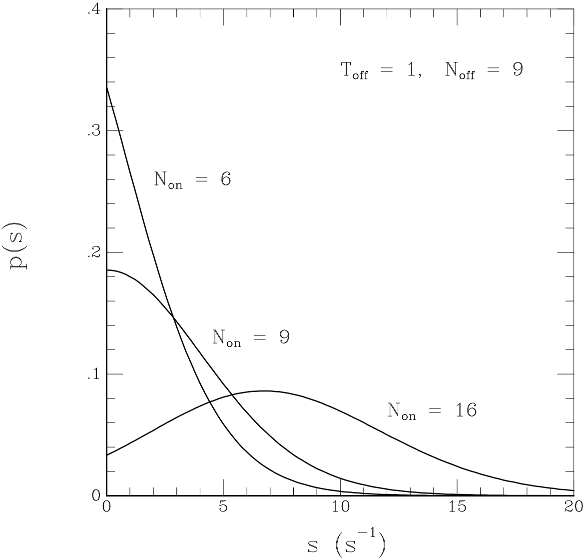
Expand $(s+b)^{N_{\rm on}}$ and do the resulting Γ integrals:

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

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)

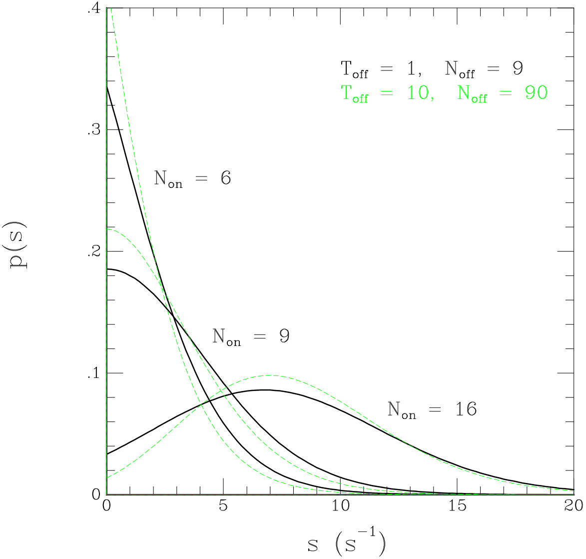
Example On/Off Posteriors—Short Integrations

$T_{\rm on} = 1$



Example On/Off Posteriors—Long Background Integrations

$T_{\text{on}} = 1$



Supplement: Two more solutions of on/off problem (including data augmentation); multibin case

Recap of Key Ideas From Examples

- Sufficient statistic: Model-dependent summary of data
- Default priors: proper, improper, symmetry, prediction, reference, minimum sample size
- Conjugate prior/likelihood pairs:
  - Beta-binomial
  - Normal-normal
  - Gamma-Poisson
- Marginalization: Generalizes background subtraction (*don't just subract!*), propagation of errors, data augmentation
- Likelihood principle
- Notable results: Bernoulli/binomial Bayes factor, Student's t, Poisson on/off, Bretthorst algorithm

Recommended exercises

- Do the flat-prior normal & Poisson calculations with *proper* priors (use the error function or the normal CDF, $\Phi(x)$ for the normal case, incomplete gamma function for Poisson case)
- Do the algebra for the normal-normal case, deriving the equations for $\widetilde{\mu}$, $\widetilde{w}$
- Show that a prior $\propto 1/r$ is a flat prior for $\lambda = \log r$
- Work through the marginalization of σ giving the Student's t distribution (see Supp)
- Work through the algebra/calculus for background marginalization:
  - *Normal case:* Complete the square in b & do Gaussian integral; complete the square in s in final result
  - *Poisson case:* Derive the C_i formula; also data augmentation version (Supp)
- Learn about the Bretthorst algorithm (GLB's book, TL's Bayesian harmonic analysis)