Introduction to Bayesian Inference

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

CASSt Summer School — June 11, 2008
1. The Big Picture
2. Foundations—Logic & Probability Theory
3. Inference With Parametric Models
   - Parameter Estimation
   - Model Uncertainty
4. Simple Examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
5. Measurement Error Applications
6. Bayesian Computation
7. Probability & Frequency
8. Endnotes: Hotspots, tools, reflections
Outline

1 The Big Picture

2 Foundations—Logic & Probability Theory

3 Inference With Parametric Models
   Parameter Estimation
   Model Uncertainty

4 Simple Examples
   Binary Outcomes
   Normal Distribution
   Poisson Distribution

5 Measurement Error Applications

6 Bayesian Computation

7 Probability & Frequency

8 Endnotes: Hotspots, tools, reflections
**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 ≡ Collection of statements comprising an act of reasoning from *premises* to a *conclusion*

A key goal of science: Explain or predict *quantitative measurements* (data!)

Data analysis constructs and appraises arguments that reason from data to interesting scientific conclusions (explanations, predictions)
The Role of Data

*Data do not speak for themselves!*

We don’t just *tabulate* data, we *analyze* data.

We gather data so they may speak for or against existing hypotheses, and guide the formation of new hypotheses.

A key role of data in science is to be among the premises in scientific arguments.
Statistical inference is but one of several interacting modes of analyzing data.
Bayesian Statistical Inference

- A different approach to *all* statistical inference problems (i.e., not just another method in the list: BLUE, maximum likelihood, $\chi^2$ 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*
Frequentist vs. Bayesian Statements

“I find conclusion $C$ based on data $D_{obs}$ . . . ”

Frequentist assessment

“It was found with a procedure that’s right 95% of the time over the set $\{D_{hyp}\}$ that includes $D_{obs}$.”

Probabilities are properties of procedures, not of particular results.

Bayesian assessment

“The strength of the chain of reasoning from $D_{obs}$ to $C$ is 0.95, on a scale where 1 = certainty.”

Probabilities are properties of specific results.

Performance must be separately evaluated (and is typically good by frequentist criteria).
Outline

1 The Big Picture
2 Foundations—Logic & Probability Theory
3 Inference With Parametric Models
   Parameter Estimation
   Model Uncertainty
4 Simple Examples
   Binary Outcomes
   Normal Distribution
   Poisson Distribution
5 Measurement Error Applications
6 Bayesian Computation
7 Probability & Frequency
8 Endnotes: Hotspots, tools, reflections
Logic—Some Essentials

“Logic can be defined as *the analysis and appraisal of arguments*”
—Gensler, *Intro to Logic*

Build arguments with propositions and logical operators/connectives

- **Propositions:** Statements that may be true or false
  
  \( \mathcal{P} : \) Universe can be modeled with ΛCDM
  
  \( A : \) \( \Omega_{\text{tot}} \in [0.9, 1.1] \)
  
  \( B : \) \( \Omega_{\Lambda} \) is not 0
  
  \( \overline{B} : \) “*not* \( B \),” i.e., \( \Omega_{\Lambda} = 0 \)

- **Connectives:**
  
  \( A \land B : \) *A and B are both* true
  
  \( A \lor B : \) *A or B is true, or both are*
Arguments

Argument: Assertion that an hypothesized conclusion, $H$, follows from premises, $\mathcal{P} = \{A, B, C, \ldots\}$ (take “,” = “and”)

Notation:

$$H \mid \mathcal{P} : \quad \text{Premises } \mathcal{P} \text{ imply } H$$

$H$ may be deduced from $\mathcal{P}$

$H$ follows from $\mathcal{P}$

$H$ is true given that $\mathcal{P}$ is true

Arguments are (compound) propositions.

Central role of arguments $\rightarrow$ special terminology for true/false:

- A true argument is valid
- A false argument is invalid or fallacious
Valid vs. Sound Arguments

**Content vs. form**

- An argument is *factually correct* iff all of its *premises are true* (it has “good content”).

- An argument is *valid* iff its conclusion *follows from* its premises (it has “good form”).

- An argument is *sound* iff it is both *factually correct and valid* (it has good form and content).

We want to make *sound* arguments. Formal logic and probability theory address validity, but there is no formal approach for addressing factual correctness → there is always a subjective element to an argument.
Factual Correctness

Although logic can teach us something about validity and invalidity, it can teach us very little about factual correctness. The question of the truth or falsity of individual statements is primarily the subject matter of the sciences.

— Hardegree, *Symbolic Logic*

To test the truth or falsehood of premisses is the task of science. . . . But as a matter of fact we are interested in, and must often depend upon, the correctness of arguments whose premisses are not known to be true.

— Copi, *Introduction to Logic*
Premises

- **Facts** — Things known to be true, e.g. *observed data*

- “**Obvious**” assumptions — Axioms, postulates, e.g., Euclid’s first 4 postulates (line segment b/t 2 points; congruency of right angles . . . )

- “**Reasonable**” or “**working**” assumptions — E.g., Euclid’s fifth postulate (parallel lines)

- *Desperate presumption!*

- Conclusions from other arguments
Deductive and Inductive Inference

**Deduction—Syllogism as prototype**

Premise 1: \( A \) implies \( H \)
Premise 2: \( A \) is true
Deduction: \( \therefore H \) is true
\( H \mid \mathcal{P} \) is valid

**Induction—Analogy as prototype**

Premise 1: \( A, B, C, D, E \) all share properties \( x, y, z \)
Premise 2: \( F \) has properties \( x, y \)
Induction: \( F \) has property \( z \)
“\( F \) has \( z \)” \( \mid \mathcal{P} \) is not strictly valid, but may still be rational (likely, plausible, probable); some such arguments are stronger than others

Boolean algebra (and/or/not over \( \{0, 1\} \)) quantifies deduction.

Bayesian probability theory (and/or/not over \( [0, 1] \)) generalizes this to quantify the strength of inductive arguments.
Real Number Representation of Induction

\[ P(H|\mathcal{P}) \equiv \text{strength of argument } H|\mathcal{P} \]

- \[ P = 0 \rightarrow \text{Argument is invalid} \]
- \[ P = 1 \rightarrow \text{Argument is valid} \]
- \[ P \in (0, 1) \rightarrow \text{Degree of deducibility} \]

A mathematical model for induction:

- ‘AND’ (product rule):
  \[ P(A \land B|\mathcal{P}) = P(A|\mathcal{P}) P(B|A \land \mathcal{P}) = P(B|\mathcal{P}) P(A|B \land \mathcal{P}) \]

- ‘OR’ (sum rule):
  \[ P(A \lor B|\mathcal{P}) = P(A|\mathcal{P}) + P(B|\mathcal{P}) - P(A \land B|\mathcal{P}) \]

- ‘NOT’:
  \[ P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P}) \]

Bayesian inference explores the implications of this model.
Interpreting Bayesian Probabilities

If we like there is no harm in saying that a probability expresses a degree of reasonable belief. . . . ‘Degree of confirmation’ has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. Essentially the notion can only be described by reference to instances where it is used. It is intended to express a kind of relation between data and consequence that habitually arises in science and in everyday life, and the reader should be able to recognize the relation from examples of the circumstances when it arises.

— Sir Harold Jeffreys, *Scientific Inference*
Physics uses words drawn from ordinary language—mass, weight, momentum, force, temperature, heat, etc.—but their technical meaning is more abstract than their colloquial meaning. We can map between the colloquial and abstract meanings associated with specific values by using specific instances as “calibrators.”

### A Thermal Analogy

<table>
<thead>
<tr>
<th>Intuitive notion</th>
<th>Quantification</th>
<th>Calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot, cold</td>
<td>Temperature, $T$</td>
<td>Cold as ice = 273K Boiling hot = 373K</td>
</tr>
<tr>
<td>uncertainty</td>
<td>Probability, $P$</td>
<td>Certainty = 0, 1 $p = 1/36$: plausible as “snake’s eyes” $p = 1/1024$: plausible as 10 heads</td>
</tr>
</tbody>
</table>
A Bit More On Interpretation

**Bayesian**

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

![Graph showing probability distribution]

**Frequentist**

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

![Graph showing frequency distribution]
Arguments Relating 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: $H_i \mid D_{\text{obs}}, I$. ($I =$ “background information,” everything deemed relevant besides the observed data)

$P(H_i \mid D_{\text{obs}}, I)$ measures the degree to which $(D_{\text{obs}}, I)$ allow one to deduce $H_i$. It provides an ordering among arguments for various $H_i$ that share common premises.

Probability theory tells us how to analyze and appraise the argument, i.e., how to calculate $P(H_i \mid D_{\text{obs}}, I)$ from simpler, hopefully more accessible probabilities.
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): $P(H_1 \lor H_2|I) = P(H_1|I) + P(H_2|I) - P(H_1, H_2|I)$

‘AND’ (product rule): $P(H_1, D|I) = P(H_1|I)P(D|H_1, I)$

$= P(D|I)P(H_1|D, I)$

‘NOT’: $P(\overline{H_1}|I) = 1 - P(H_1|I)$
Bayes’s Theorem (BT)

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

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

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

Solve for the posterior probability:

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

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

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

norm. const. $P(D_{obs}|I) = \text{prior predictive}$
Law of Total Probability (LTP)

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

\[
\sum_i P(A, B_i|I) = \sum_i P(B_i|A, I)P(A|I) = P(A|I)
\]

\[
= \sum_i P(B_i|I)P(A|B_i, I)
\]

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 \( A = D_{\text{obs}} , B_i = H_i \); then

\[
P(D_{\text{obs}} | I) = \sum_i P(D_{\text{obs}}, H_i | I) \\
= \sum_i P(H_i | I) P(D_{\text{obs}} | H_i, I)
\]

prior predictive for \( D_{\text{obs}} \) = Average likelihood for \( H_i \)
(a.k.a. marginal likelihood)

**Normalization**

For exclusive, exhaustive \( H_i \),

\[
\sum_i P(H_i | \cdots) = 1
\]
Well-Posed Problems

The rules express desired probabilities in terms of other probabilities.

To get a numerical value *out*, at some point we have to put numerical values *in*.

*Direct probabilities* are probabilities with numerical values determined directly by premises (via modeling assumptions, symmetry arguments, previous calculations, desperate presumption . . . ).

An inference problem is *well posed* only if all the needed probabilities are assignable based on the premises. We may need to add new assumptions as we see what needs to be assigned. We may not be entirely comfortable with what we need to assume! (Remember Euclid’s fifth postulate!)

Should explore how results depend on uncomfortable assumptions ("robustness").
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.
- LTP: If the premises allow multiple arguments for a hypothesis, its appraisal must account for all of them.
Outline

1. The Big Picture
2. Foundations—Logic & Probability Theory
3. Inference With Parametric Models
   - Parameter Estimation
   - Model Uncertainty
4. Simple Examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
5. Measurement Error Applications
6. Bayesian Computation
7. Probability & Frequency
8. Endnotes: Hotspots, tools, reflections
Inference With Parametric Models

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

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

The $\theta_i$ dependence when we fix attention on the observed data is the likelihood function:

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

We may be uncertain about $i$ (model uncertainty) or $\theta_i$ (parameter uncertainty).
Three Classes of Problems

Parameter Estimation

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

Model Assessment

• Model comparison: Premise = $\{M_i\}$
$\rightarrow$ What can we say about $i$?

• Model adequacy/GoF: Premise = $M_1 \lor$ “all” alternatives
$\rightarrow$ Is $M_1$ adequate?

Model Averaging

Models share some common params: $\theta_i = \{\phi, \eta_i\}$
$\rightarrow$ What can we say about $\phi$ w/o committing to one model?
(Systematic error is an example)
Parameter Estimation

Problem statement

\[ I = \text{Model } M \text{ with parameters } \theta \text{ (} + \text{ any add’l info) } \]

\[ H_i = \text{statements about } \theta; \text{ e.g. } “\theta \in [2.5, 3.5],” \text{ or } “\theta > 0” \]

Probability for any such statement can be found using a probability density function (PDF) for \( \theta \):

\[
P(\theta \in [\theta, \theta + d\theta] | \cdots) = f(\theta) d\theta
\]

\[
= p(\theta | \cdots) d\theta
\]

Posterior probability density

\[
p(\theta | D, M) = \frac{p(\theta | M) L(\theta)}{\int d\theta p(\theta | M) L(\theta)}
\]
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 $\Delta$ 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 $\psi$, nuisance parameters $\phi$
  - Marginal dist’n for $\psi$: $p(\psi|D, M) = \int d\phi p(\psi, \phi|D, M)$
Nuisance Parameters and Marginalization

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

*Example*

We have data from measuring a rate \( r = s + b \) 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

\[ p(s|D, M) = \int db \ p(s, b|D, M) \]
\[ \propto p(s|M) \int db \ p(b|s) \mathcal{L}(s, b) \]
\[ \equiv p(s|M)\mathcal{L}_m(s) \]

with \( \mathcal{L}_m(s) \) the marginal likelihood for \( s \). For broad prior,

\[ \mathcal{L}_m(s) \approx p(\hat{b}_s|s) \mathcal{L}(s, \hat{b}_s) \delta b_s \]

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

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

Posterior probability for a model

\[
p(M_i|D, I) = \frac{p(M_i|I) \frac{p(D|M_i, I)}{p(D|I)}}{p(M_i|I) \mathcal{L}(M_i)} \propto p(M_i|I) \mathcal{L}(M_i)
\]

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

A ratio of probabilities for two propositions using the same premises is called the **odds** favoring one over the other:

\[
O_{ij} \equiv \frac{p(M_i|D, I)}{p(M_j|D, I)} = \frac{p(M_i|I)}{p(M_j|I)} \times \frac{p(D|M_j, I)}{p(D|M_j, I)}
\]

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.
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

\[ p(D|M_i) = \int d\theta_i \ p(\theta_i|M) \ L(\theta_i) \approx p(\hat{\theta}_i|M)L(\hat{\theta}_i)\delta\theta_i \]

\[ \approx L(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \]

\[ = \text{Maximum Likelihood} \times \text{Occam Factor} \]

Models with more parameters often make the data more probable — for the best fit
Occam factor penalizes models for “wasted” volume of parameter space
Quantifies intuition that models shouldn’t require fine-tuning
Model Averaging

Problem statement

\[ I = (M_1 \vee M_2 \vee \ldots) \] — Specify a set of models
Models all share a set of “interesting” parameters, \( \phi \)
Each has different set of nuisance parameters \( \eta_i \) (or different prior info about them)
\( H_i = \) statements about \( \phi \)

Model averaging

Calculate posterior PDF for \( \phi \):

\[
p(\phi|D, I) = \sum_i p(M_i|D, I) p(\phi|D, M_i)
\]

\[
\propto \sum_i \mathcal{L}(M_i) \int d\eta_i p(\phi, \eta_i|D, M_i)
\]

The model choice is a (discrete) nuisance parameter here.
Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over sample space & typically optimize over parameter 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”).
Roles of the Prior

Prior has two roles

- Incorporate any relevant prior information
- Convert likelihood from “intensity” to “measure”
  → Accounts for size of hypothesis space

Physical analogy

Heat: \[ Q = \int dV \, c_v(r) \, T(r) \]

Probability: \[ P \propto \int d\theta \, p(\theta|I) \mathcal{L}(\theta) \]

Maximum likelihood focuses on the “hottest” hypotheses. Bayes focuses on the hypotheses with the most “heat.”

A high-\( T \) region may contain little heat if its \( c_v \) is low or if its volume is small.

A high-\( \mathcal{L} \) region may contain little probability if its prior is low or if its volume is small.
Recap of 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
Outline

1. The Big Picture
2. Foundations—Logic & Probability Theory
3. Inference With Parametric Models
   - Parameter Estimation
   - Model Uncertainty
4. Simple Examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
5. Measurement Error Applications
6. Bayesian Computation
7. Probability & Frequency
8. Endnotes: Hotspots, tools, reflections
Binary Outcomes: Parameter Estimation

$M = \text{Existence of two outcomes, } S \text{ and } F; \text{ each trial has same probability for } S \text{ or } F$

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

$D = \text{Sequence of results from } N \text{ observed trials:}$

FFSSSSSFSSSFS ($n = 8 \text{ successes in } N = 12 \text{ trials}$)

Likelihood:

$$p(D|\alpha, M) = p(\text{failure}|\alpha, M) \times p(\text{success}|\alpha, M) \times \cdots$$

$$= \alpha^n (1 - \alpha)^{N-n}$$

$$= \mathcal{L}(\alpha)$$
Prior

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

- Intuition: Don’t prefer any $\alpha$ 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:

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

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

\[ p(D|M) = \int d\alpha \ \alpha^n (1 - \alpha)^{N-n} \]

\[ = B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!} \]

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**.
Binary Outcomes: Model Comparison

Equal Probabilities?

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

Maximum Likelihoods

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

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

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

Maximum likelihoods favor \( M_2 \) (failures more probable).
Bayes Factor (ratio of model likelihoods)

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

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

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 \( S = \) a sequence of flips with \( n \) heads.

\[
p(n|\alpha, M) = \sum_S p(S|\alpha, M) p(n|S, \alpha, M)
\]

\[
= \alpha^n (1 - \alpha)^{N-n} C_{n,N}
\]

\( C_{n,N} = \# \) 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 \( \alpha, N \).
\[ p(\alpha|n, M) = \frac{\frac{N!}{n!(N-n)!} \alpha^n (1 - \alpha)^{N-n}}{p(n|M)} \]

\[ p(n|M) = \frac{\frac{N!}{n!(N-n)!}}{\int d\alpha \, \alpha^n (1 - \alpha)^{N-n}} = \frac{1}{N + 1} \]

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

*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 predefined 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:

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

$$= \frac{(N - 1)!}{(n - 1)!(N - n)!} \alpha^{n} (1 - \alpha)^{N-n}$$

The *negative binomial distribution* for $N$ given $\alpha$, $n$. 
Posterior

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

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

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

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, $W(N)$:

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

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 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.

Example: Predict 10% of sample is Type A; observe $n_A = 5$ for $N = 96$
Significance test accepts $\alpha = 0.1$ for binomial sampling;
$p(>\chi^2|\alpha = 0.1) = 0.12$
Significance test rejects $\alpha = 0.1$ for negative binomial sampling;
$p(>\chi^2|\alpha = 0.1) = 0.03$
Inference With Normals/Gaussians

Gaussian PDF

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

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

Parameters

\[ \mu = \langle x \rangle \equiv \int dx \, x \, p(x|\mu, \sigma) \]
\[ \sigma^2 = \langle (x - \mu)^2 \rangle \equiv \int dx \, (x - \mu)^2 \, p(x|\mu, \sigma) \]
Gauss’s Observation: Sufficiency

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

\[
p(D|\mu, \sigma, M) = \prod_i p(d_i|\mu, \sigma, M)
= \prod_i p(\epsilon_i = d_i - \mu|\mu, \sigma, M)
= \prod_i \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(d_i - \mu)^2}{2\sigma^2} \right]
= \frac{1}{\sigma^N (2\pi)^{N/2}} e^{-Q(\mu)/2\sigma^2}
\]
Find dependence of $Q$ on $\mu$ by completing the square:

$$Q = \sum_i (d_i - \mu)^2$$

$$= \sum_i d_i^2 + N\mu^2 - 2N\mu \bar{d} \quad \text{where} \quad \bar{d} \equiv \frac{1}{N} \sum_i d_i$$

$$= N(\mu - \bar{d})^2 + Nr^2 \quad \text{where} \quad r^2 \equiv \frac{1}{N} \sum_i (d_i - \bar{d})^2$$

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

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

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!
Problem specification

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

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

Likelihood

\[
p(D|\mu, \sigma, M) = \frac{1}{\sigma^N(2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \
\propto \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right)
\]
“Uninformative” prior

Translation invariance \( \Rightarrow p(\mu) \propto C \), a constant.
This prior is improper unless bounded.

Prior predictive/normalization

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

\[\ldots\text{ minus a tiny bit from tails, using a proper prior.}\]
**Posterior**

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

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

68.3% HPD credible region for \( \mu \) is \( \bar{d} \pm \sigma/\sqrt{N} \).

Note that \( C \) drops out \( \rightarrow \) limit of infinite prior range is well behaved.
Informative Conjugate Prior

Use a normal prior, \( \mu \sim N(\mu_0, w^2_0) \)

Posterior

Normal \( N(\tilde{\mu}, \tilde{w}^2) \), but mean, std. deviation “shrink” towards prior. Define \( B = \frac{w^2}{w^2 + w^2_0} \), so \( B < 1 \) and \( B = 0 \) when \( w_0 \) is large. Then

\[
\begin{align*}
\tilde{\mu} &= (1 - B) \cdot \bar{d} + B \cdot \mu_0 \\
\tilde{w} &= w \cdot \sqrt{1 - B}
\end{align*}
\]

“Principle of stable estimation:” The prior affects estimates only when data are not informative relative to prior.
Estimating a Normal Mean: Unknown $\sigma$

**Problem specification**

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

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

**Likelihood**

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

where $Q = N \left[ r^2 + (\mu - \bar{d})^2 \right]$
Uninformative Priors

Assume priors for $\mu$ and $\sigma$ 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 $\mu$, $\sigma$

$$ p(\mu, \sigma|D, M) \propto \frac{1}{\sigma^{N+1}} e^{-Q(\mu)/2\sigma^2} $$
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 \( |d\sigma| = \tau^{-3/2} \sqrt{\frac{Q}{2}} \)

\[ \Rightarrow p(\mu|D, M) \propto 2^{N/2} Q^{-N/2} \int d\tau \frac{\tau^{N}}{2}^{-1} e^{-\tau} \]

\[ \propto Q^{-N/2} \]
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

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

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 $p(n|M) \sim \text{Const}$:
  \[ p(r|M) = \frac{1}{r_u} \]
Prior predictive

\[
p(n|M) = \frac{1}{r_u} \frac{1}{n!} \int_0^{r_u} dr (rT)^n e^{-rT}
\]

\[
= \frac{1}{r_u T} \frac{1}{n!} \int_0^{r_u T} d(rT) (rT)^n e^{-rT}
\]

\[
\approx \frac{1}{r_u T} \quad \text{for} \quad r_u \gg \frac{n}{T}
\]

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 \( \alpha \) 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:

\[
E(x) = s\nu \quad \text{Var}(x) = s^2\nu
\]

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)
$T = 2 \, \text{s}, \, n = 10$
The flat prior

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

Require (discrete) predictive distribution to be flat:

$$p(n|I) = \int dr \ p(r|I)p(n|r, I) = C$$

$$\rightarrow p(r|I) = C$$

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 $N_{off}$ photons in interval $T_{off}$

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

- Infer $s$

Conventional solution

\[
\hat{b} = \frac{N_{off}}{T_{off}}; \quad \sigma_b = \sqrt{\frac{N_{off}}{T_{off}}}
\]
\[
\hat{r} = \frac{N_{on}}{T_{on}}; \quad \sigma_r = \sqrt{\frac{N_{on}}{T_{on}}}
\]
\[
\hat{s} = \hat{r} - \hat{b}; \quad \sigma_s = \sqrt{\sigma^2_r + \sigma^2_b}
\]

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

Spectra of X-Ray Sources

Bassani et al. 1989

Di Salvo et al. 2001

CYGNOUS X-1

non-flaring

Energy (keV)

Photons cm$^{-2}$ s$^{-1}$

10$^{-4}$ 10$^{-3}$ 10$^{-2}$ 10$^{-1}$ 0.1 1

Gamma-ray Flux (Ph/cm$^2$ s sec)

0 0.5 1.0 1.5 2.0 2.5 3.0

Hard X-Ray Flux (Ph/cm$^2$ s sec)

x10^{-2} 5 10 15 20 25

(A) (D) (E) (G) (H) (I) (B) (C)
Spectrum of Ultrahigh-Energy Cosmic Rays

Nagano & Watson 2000

\[ \log_{10}(E) \text{ (eV)} \]

\[ \text{Flux} \times 10^{24} \text{ (eV}^2\text{m}^{-2}\text{s}^{-1}\text{sr}^{-1}) \]

AGASA
HiRes-1 Monocular
HiRes-2 Monocular

Knee
2nd Knee
Ankle
speculated GZK cutoff

EAS experiments

HiRes Team 2007
“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)
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 $b$):

$$p(s|D, M) \propto \exp \left[ - \frac{(s - \hat{s})^2}{2\sigma^2_s} \right] \quad \hat{s} = \hat{r} - \hat{b} \quad \sigma^2_s = \sigma^2_r + \sigma^2_b$$

⇒ Background subtraction is a special case of background marginalization.
Bayesian Solution to On/Off Problem

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

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

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

$$
p(s, b|N_{\text{on}}) \propto p(s)p(b)[(s + b)T_{\text{on}}]^{N_{\text{on}}} e^{-(s+b)T_{\text{on}}} \parallel I_{\text{all}}
$$

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

$$
p(s, b|N_{\text{on}}, I_{\text{all}}) \propto (s + b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-sT_{\text{on}}} e^{-b(T_{\text{on}}+T_{\text{off}})}
$$
Now marginalize over $b$;

$$p(s|N_{on}, I_{all}) = \int db \, p(s, b \mid N_{on}, I_{all})$$

$$\propto \int db \, (s + b)^{N_{on}} b^{N_{off}} e^{-sT_{on}} e^{-b(T_{on} + T_{off})}$$

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

$$p(s|N_{on}, I_{all}) = \sum_{i=0}^{N_{on}} C_i \frac{T_{on} (sT_{on})^i e^{-sT_{on}}}{i!}$$

$$C_i \propto \left(1 + \frac{T_{off}}{T_{on}}\right)^i \frac{(N_{on} + N_{off} - i)!}{(N_{on} - i)!}$$

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

\[ T_{on} = 1 \]

\[ T_{off} = 1, \quad N_{off} = 9 \]

\[ N_{on} = 6 \]

\[ N_{on} = 9 \]

\[ N_{on} = 16 \]
Example On/Off Posteriors—Long Background Integrations

$T_{on} = 1$

$T_{off} = 1, N_{off} = 9$

$T_{off} = 10, N_{off} = 90$

$p(s)$

$N_{on} = 6$

$N_{on} = 9$

$N_{on} = 16$

$s \ (s^{-1})$
Outline

1. The Big Picture
2. Foundations—Logic & Probability Theory
3. Inference With Parametric Models
   - Parameter Estimation
   - Model Uncertainty
4. Simple Examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
5. Measurement Error Applications
6. Bayesian Computation
7. Probability & Frequency
8. Endnotes: Hotspots, tools, reflections
Empirical Number Counts Distributions

*Star counts, galaxy counts, GRBs, TNOs* ...
Selection Effects and Measurement Error

- 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 (Nov 2007)

QSO hardness vs. luminosity (Kelly 2007)
Eddington, Jeffreys (1920s – 1940)

Malmquist, Lutz-Kelker

- Joint accounting for truncation and (intrinsic) scatter in 2-D data (flux + distance indicator, parallax)
- Assume homogeneous spatial distribution
Many rediscoveries of “scatter biases”

- Radio sources (1970s)
- Galaxies (Eddington, Malmquist; 1990s)
- Linear regression (1990s)
- GRBs (1990s)
- X-ray sources (1990s; 2000s)
- TNOs/KBOs (c. 2000)
- Galaxy redshift dist’ns (2007+)
- …
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 $\theta$ from

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

\[ \mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta) \]
But what if the $x$ data are noisy, $D_i = \{x_i + \epsilon_i\}$?

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

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

*Marginalize* (sum probabilities) over $\{x_i\}$ to summarize for $\theta$.

*Marginalize* over $\theta$ to summarize results for $\{x_i\}$.

Key point: *Maximizing over $x_i$ and integrating over $x_i$ can give very different results!*
A two-level *multi-level model* (MLM).

\[
\mathcal{L}(\theta, \{x_i\}) \equiv p(\{D_i\}|\theta, \{x_i\}) = \prod_i p(D_i|x_i)f(x_i|\theta) = \prod_i \ell_i(x_i)f(x_i|\theta)
\]
Example—Distribution of Source Fluxes

Measure $m = -2.5 \log(\text{flux})$ from sources following a “rolling power law” distribution (inspired by trans-Neptunian objects)

$$f(m) \propto 10^{\left[\alpha(m-23) + \alpha'(m-23)^2\right]}$$

Simulate 100 surveys of populations drawn from the same dist’n. Simulate data for photon-counting instrument, fixed count threshold. Measurements have uncertainties 1% (bright) to $\approx 30\%$ (dim).

Analyze simulated data with maximum (“profile”) likelihood and Bayes.
Parameter estimates from Bayes (circles) and maximum likelihood (crosses):

\[ N = 100 \]

\[ N = 1000 \]

Uncertainties don’t average out!
Smaller measurement error only postpones the inevitable:

Similar toy survey, with parameters to mimic SDSS QSO surveys (few % errors at dim end):
Bayesian MLMs in Astronomy

- **Directional & spatio-temporal coincidences:**
  - GRB repetition (Luo$^+$ 1996; Graziani$^+$ 1996)
  - GRB host ID (Band 1998; Graziani$^+$ 1999)
  - VO cross-matching (Badavi’ari & Szalay 2008)

- **Magnitude surveys/number counts/“log N–log S”:**
  - GRB peak flux dist’n (Loredo & Wasserman 1998);
  - TNO/KBO magnitude distribution (Gladman$^+$ 1998; Petit$^+$ 2008)

- **Dynamic spectroscopy:** SN 1987A neutrinos, uncertain energy vs. time (Loredo & Lamb 2002)

- **Linear regression:** QSO hardness vs. luminosity (Kelly 2007)
Outline

1. The Big Picture
2. Foundations—Logic & Probability Theory
3. Inference With Parametric Models
   - Parameter Estimation
   - Model Uncertainty
4. Simple Examples
   - Binary Outcomes
   - Normal Distribution
   - Poisson Distribution
5. Measurement Error Applications
6. Bayesian Computation
7. Probability & Frequency
8. Endnotes: Hotspots, tools, reflections
Bayesian Computation

**Large sample size: Laplace approximation**

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

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

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

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

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

1  The Big Picture
2  Foundations—Logic & Probability Theory
3  Inference With Parametric Models
   Parameter Estimation
   Model Uncertainty
4  Simple Examples
   Binary Outcomes
   Normal Distribution
   Poisson Distribution
5  Measurement Error Applications
6  Bayesian Computation
7  Probability & Frequency
8  Endnotes: Hotspots, tools, reflections
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
- Examples of Bayesian/frequentist differences
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” → First use of Bayes’s theorem:
Probability for success in next trial of i.i.d. sequence:

$$E\alpha \to \frac{N_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \to \infty$$
Subtle Relationships For Non-IID Cases

*Predict frequency in dependent trials*

\[ r_t = \text{result of trial } t; \ p(r_1, r_2 \ldots r_N | M) \text{ known}; \text{ predict } f: \]

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

where

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

*Expected frequency of outcome in many trials = average probability for outcome across trials.*

*But* also find that \( \sigma_f \) needn’t converge to 0.

**Infer probabilities for different but 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.)
Outline

1 The Big Picture
2 Foundations—Logic & Probability Theory
3 Inference With Parametric Models
   Parameter Estimation
   Model Uncertainty
4 Simple Examples
   Binary Outcomes
   Normal Distribution
   Poisson Distribution
5 Measurement Error Applications
6 Bayesian Computation
7 Probability & Frequency
8 Endnotes: Hotspots, tools, reflections
Some Bayesian Astrostatistics Hotspots

- **Cosmology**
  - Parametric modeling of CMB, LSS, SNe Ia $\rightarrow$ cosmo params
  - Nonparametric modeling of SN Ia multicolor light curves
  - Nonparametric “emulation” of cosmological models

- **Extrasolar planets**
  - Parametric modeling of Keplerian reflex motion (planet detection, orbit estimation)
  - Optimal scheduling via Bayesian experimental design

- **Photon counting data (X-rays, $\gamma$-rays, cosmic rays)**
  - Upper limits, hardness ratios
  - Parametric spectroscopy (line detection, etc.)

- **Gravitational wave astronomy**
  - Parametric modeling of binary inspirals
  - Hi-multiplicity parametric modeling of white dwarf background
Tools for Computational Bayes

Astronomer/Physicist Tools

- **BIE** [http://www.astro.umass.edu/~weinberg/proto_bie/](http://www.astro.umass.edu/~weinberg/proto_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.

- **XSpec, CIAO/Sherpa**
  Both environments have some basic Bayesian capability (including basic MCMC in XSpec)

- **CosmoMC** [http://cosmologist.info/cosmomc/](http://cosmologist.info/cosmomc/)
  Parameter estimation for cosmological models using CMB and other data via MCMC

- **ExoFit** [http://zuserver2.star.ucl.ac.uk/~lahav/exofit.html](http://zuserver2.star.ucl.ac.uk/~lahav/exofit.html)
  Adaptive MCMC for fitting exoplanet RV data

- **CDF Bayesian Limit Software**
  Limits for Poisson counting processes, with background & efficiency uncertainties

- **root** [http://root.cern.ch/](http://root.cern.ch/)
  Bayesian support? (BayesDivide)

  Several self-contained Bayesian modules (Gaussian, Poisson, directional processes); Parametric Inference Engine (PIE) supports $\chi^2$, likelihood, and Bayes
**Python**

- **PyMC** [http://trichech.us/pymc](http://trichech.us/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/](http://simpy.sourceforge.net/)
  SimPy (rhymes with "Blimpie") is a process-oriented public-domain package for discrete-event simulation.

- **RSPython** [http://www.omegahat.org/](http://www.omegahat.org/)
  Bi-directional communication between Python and R

  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/](http://www.ailab.si/orange/)
  Component-based data mining, with preprocessing, modeling, and exploration components. Python/GUI interfaces to C++ implementations. Some Bayesian components.

  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/src/contrib/Views/Bayesian.html](http://cran.r-project.org/src/contrib/Views/Bayesian.html)
  Overview of many R packages implementing various Bayesian models and methods

  RPython, RMatlab, R-Xlisp

- **BOA** [http://www.public-health.uiowa.edu/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](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
Java

  Java environment for statistical computing, being developed by XLisp-stat and R developers

- **Hydra** [http://research.warnes.net/projects/mcmc/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](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)
C/C++/Fortran

- **BayeSys 3** [http://www.inference.phy.cam.ac.uk/bayesys/](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](http://www.cs.utoronto.ca/~radford/fbm.software.html)
  Flexible Bayesian Modeling: MCMC for simple Bayes, 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](http://www.sci.wsu.edu/math/faculty/genz/homepage)
  Adaptive quadrature, randomized quadrature, Monte Carlo integration

- **BIE, CDF Bayesian limits** (see above)
Other Statisticians’ & Engineers’ Tools

- **BUGS/WinBUGS** [http://www.mrc-bsu.cam.ac.uk/bugs/](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/](http://mathstat.helsinki.fi/openbugs/)
  BUGS on Windows and Linux, and from inside the R

  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/](http://choosh.csee.ogi.edu/rebel/)
  Library supporting recursive Bayesian estimation in Matlab (Kalman filter, particle filters, sequential Monte Carlo).
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’.

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.