Introduction to Bayesian inference in astronomy

Tom Loredo
Center for Radiophysics & Space Research, Cornell University
http://www.astro.cornell.edu/staff/loredo/bayes/

CASt Summer School — 4 June 2015
**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:* Constructing and appraising arguments that reason from data to interesting scientific conclusions (explanations, predictions)
The role of data

Data do not speak for themselves!

“No body of data tells us all we need to know about its own analysis.”
— John Tukey, EDA

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

• Bayesian inference uses probability theory to quantify the strength of data-based arguments (i.e., a more abstract view than restricting PT to describe variability in repeated “random” experiments)

• A different approach to all statistical inference problems (i.e., not just another method in the list: BLUE, linear regression, least squares/\(\chi^2\) minimization, maximum likelihood, ANOVA, survival analysis, LDA classification . . . )

• Focuses on deriving consequences of modeling assumptions rather than devising and calibrating procedures
Frequentist vs. Bayesian statements

“The data $D_{\text{obs}}$ support conclusion $C$ . . . ”

Frequentist assessment

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

Probabilities are properties of procedures, not of particular results.

Bayesian assessment

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

Probabilities are associated with specific, observed data. Long-run performance must be separately evaluated (and is typically good by frequentist criteria).
Agenda

1. Confidence intervals vs. credible intervals
2. Probability theory as generalized logic
3. Probability theory for data analysis: Three theorems
4. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty
5. Multilevel models for measurement error
6. Bayesian computation
Agenda

1. Confidence intervals vs. credible intervals
2. Probability theory as generalized logic
3. Probability theory for data analysis: Three theorems
4. Inference with parametric models
   Parameter Estimation
   Model Uncertainty
5. Multilevel models for measurement error
6. Bayesian computation
Problem

Estimate the location (mean) of a Gaussian distribution from a set of samples $D = \{x_i\}, i = 1$ to $N$. Report a region summarizing the uncertainty.

Model

$$p(x_i|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

Equivalently, $x_i \sim \mathcal{N}(\mu, \sigma^2)$

Here assume $\sigma$ is known; we are uncertain about $\mu$. 
Classes of variables

- $\mu$ is the unknown we seek to estimate—the parameter. The parameter space is the space of possible values of $\mu$—here the real line (perhaps bounded). Hypothesis space is a more general term.

- A particular set of $N$ data values $D = \{x_i\}$ is a sample. The sample space is the $N$-dimensional space of possible samples.

Standard inferences

Let $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

- “Standard error” (rms error) is $\sigma/\sqrt{N}$
- “$1\sigma$” interval: $\bar{x} \pm \sigma/\sqrt{N}$ with conf. level CL = 68.3%
- “$2\sigma$” interval: $\bar{x} \pm 2\sigma/\sqrt{N}$ with CL = 95.4%
Some simulated data

Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

What is the CL associated with this interval?
Some simulated data

Take \( \mu = 5 \) and \( \sigma = 4 \) and \( N = 16 \), so \( \sigma / \sqrt{N} = 1 \).

What is the CL associated with this interval?

The confidence level for this interval is 79.0\%. 

Green interval: $\bar{x} \pm 2\sigma / \sqrt{N}$

Blue interval: Let $x_{(k)} \equiv k$'th order statistic
Report $[x_{(6)}, x_{(11)}]$ (i.e., leave out 5 outermost each side)

Moral

The confidence level is a property of the procedure, not of the particular interval reported for a given dataset.
Performance of intervals

Intervals for 15 datasets
Confidence interval for a normal mean

Suppose we have a sample of \( N = 5 \) values \( x_i \),

\[
x_i \sim N(\mu, 1)
\]

We want to estimate \( \mu \), including some quantification of uncertainty in the estimate: an interval with a probability attached.

Frequentist approaches: method of moments, BLUE, least-squares/\( \chi^2 \), maximum likelihood

Focus on likelihood (equivalent to \( \chi^2 \) here); this is closest to Bayes.

\[
\mathcal{L}(\mu) = p(\{x_i\} | \mu) = \prod_i \frac{1}{\sigma \sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \quad \sigma = 1
\]

\[\propto e^{-\chi^2(\mu)/2}\]

Estimate \( \mu \) from maximum likelihood (minimum \( \chi^2 \)). Define an interval and its coverage frequency from the \( \mathcal{L}(\mu) \) curve.
Construct an interval procedure for known $\mu$

Likelihoods for 3 simulated data sets, $\mu = 0$
Likelihoods for 100 simulated data sets, $\mu = 0$
Explore dependence on $\mu$
Likelihoods for 100 simulated data sets, $\mu = 3$

Luckily the $\Delta \log \mathcal{L}$ distribution is the same!
($\Delta \log \mathcal{L}$ is a pivotal quantity)

If it weren’t, define confidence level = maximum coverage over all $\mu$ (confidence level = conservative guarantee of coverage).

**Parametric bootstrap:** Skip this step; just report the coverage based on $\mu = \hat{\mu}(\{x_i\})$ for the observed data. Theory shows the error in the coverage falls faster than $\sqrt{N}$. 
Apply to observed sample

Report the green region, with coverage as calculated for ensemble of hypothetical data (green region, previous slide).
Likelihood to probability via Bayes’s theorem

Recall the likelihood, \( \mathcal{L}(\mu) \equiv p(D_{\text{obs}}|\mu) \), is a probability for the observed data, but not for the parameter \( \mu \).

Convert likelihood to a probability distribution over \( \mu \) via Bayes’s theorem:

\[
p(A, B) = p(A)p(B|A) \\
= p(B)p(A|B) \\
\rightarrow p(A|B) = p(A) \frac{p(B|A)}{p(B)}, \quad \text{Bayes’s th.}
\]

\[
\Rightarrow p(\mu|D_{\text{obs}}) \propto \pi(\mu)\mathcal{L}(\mu)
\]

\( p(\mu|D_{\text{obs}}) \) is called the posterior probability distribution.

This requires a prior probability density, \( \pi(\mu) \), often taken to be constant over the allowed region if there is no significant information available (or sometimes constant w.r.t. some reparameterization motivated by a symmetry in the problem).
For the Gaussian example, a bit of algebra ("complete the square") gives:

\[
\mathcal{L}(\mu) \propto \prod_i \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
\propto \exp \left[ -\frac{1}{2} \sum_i \frac{(x_i - \mu)^2}{\sigma^2} \right]
\]

\[
\propto \exp \left[ -\frac{(\mu - \bar{x})^2}{2(\sigma/\sqrt{N})^2} \right]
\]

The likelihood is Gaussian in \( \mu \).
Flat prior \( \rightarrow \) posterior density for \( \mu \) is \( \mathcal{N}(\bar{x}, \sigma^2/N) \).
Bayesian credible region

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood.
Credible region via Monte Carlo: *posterior sampling*

Sample Space

Parameter Space

Joint Space

Normalized $L(\mu)$

200 post. samples
Posterior summaries

- Posterior mean is $\langle \mu \rangle \equiv \int d\mu \, \mu \, p(\mu | D_{\text{obs}}) = \bar{x}$
- Posterior mode is $\hat{\mu} = \bar{x}$
- Posterior std dev’n is $\sigma / \sqrt{N}$
- $\bar{x} \pm \sigma / \sqrt{N}$ is a 68.3% credible region:

\[
\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} d\mu \, p(\mu | D_{\text{obs}}) \approx 0.683
\]

- $\bar{x} \pm 2\sigma / \sqrt{N}$ is a 95.4% credible region

The credible regions above are highest posterior density credible regions (HPD regions). These are the smallest regions with a specified probability content.

These reproduce familiar frequentist results, but this is a coincidence due to special properties of Gaussians.
Confidence region calculation

Likelihoods for 100 simulated data sets, $\mu = 0$
When They’ll Differ

Both approaches report \( \mu \in [\bar{x} - \sigma/\sqrt{N}, \bar{x} + \sigma/\sqrt{N}] \), and assign 68.3% to this interval (with different meanings).

This matching is a \textit{coincidence}!

When might results differ? (\( F = \) frequentist, \( B = \) Bayes)

- If \( F \) procedure doesn’t use likelihood directly
- If \( F \) procedure properties depend on params (nonlinear models, need to find pivotal quantities)
- If likelihood shape varies strongly between datasets (conditional inference, ancillary statistics, recognizable subsets)
- If there are extra uninteresting parameters (nuisance parameters, adjusted profile likelihood, conditional inference)
- If \( B \) uses important prior information

Also, for a different task—comparison of parametric models—the approaches are \textit{qualitatively} different (significance tests & info criteria vs. Bayes factors)
Supplement — Multivariate confidence and credible regions: *parametric bootstrapping* vs. *posterior sampling*
Brad Efron, ASA President (2005)

The 250-year debate between Bayesians and frequentists is unusual among philosophical arguments in actually having *important practical consequences*... The physicists I talked with were really bothered by our 250 year old Bayesian-frequentist argument. Basically there’s only one way of doing physics but there seems to be at least two ways to do statistics, and they don’t always give the same answers....

Broadly speaking, Bayesian statistics dominated 19th Century statistical practice while the 20th Century was more frequentist. What’s going to happen in the 21st Century?... I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning....
Pragmatists might argue that good statisticians can get sensible answers under Bayes or frequentist paradigms; indeed maybe two philosophies are better than one, since they provide more tools for the statistician’s toolkit.... I am discomforted by this “inferential schizophrenia.” Since the Bayesian (B) and frequentist (F) philosophies can differ even on simple problems, at some point decisions seem needed as to which is right. I believe our credibility as statisticians is undermined when we cannot agree on the fundamentals of our subject....

An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that calibrated Bayes . . . captures the strengths of both approaches and provides a roadmap for future advances.

[Calibrated Bayes = Bayesian inference within a specified space of models + frequentist approaches for model checking; Andrew Gelman uses “Bayesian data analysis” similarly] (see TL’s arXiv:1208.3035 for discussion/references)
Agenda

1. Confidence intervals vs. credible intervals

2. Probability theory as generalized logic

3. Probability theory for data analysis: Three theorems

4. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

5. Multilevel models for measurement error

6. Bayesian computation
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
  
  \( P \): Universe can be modeled with \( \Lambda \)CDM
  
  \( A \): \( \Omega_{\text{tot}} \in [0.9, 1.1] \)
  
  \( B \): \( \Omega_{\Lambda} \) is not 0
  
  \( \overline{B} \): “not \( B \),” i.e., \( \Omega_{\Lambda} = 0 \)

- **Connectives**:

  \[ A \land B : \quad A \text{ and } B \text{ are both true} \]
  
  \[ A \lor B : \quad A \text{ or } B \text{ is true, or both are} \]
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).

Deductive logic (and probability theory) addresses *validity*.

We want to make *sound* arguments. There is no formal approach for addressing factual correctness → there is always a subjective element to an argument.
Deductive and inductive inference

Deduction—Syllogism as prototype

Premise 1: $A$ implies $H$
Premise 2: $A$ is true
Deduction: $\therefore H$ is true

$H|\mathcal{P}$ 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 $z$” $|\mathcal{P}$ 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.
Representing induction with $[0, 1]$ calculus

$P(H|\mathcal{P}) \equiv$ strength of argument $H|\mathcal{P}$

- $P = 1 \rightarrow$ Argument is *deductively valid*
- $P = 0 \rightarrow$ Premises imply $\overline{H}$
- $\in (0, 1) \rightarrow$ Degree of deducibility

*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’:

$P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$
Pierre Simon Laplace (1819)

Probability theory is nothing but *common sense reduced to calculation.*

James Clerk Maxwell (1850)

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic, but the actual science of *Logic is conversant at present only with things either certain, impossible, or entirely doubtful,* none of which (fortunately) we have to reason on. Therefore *the true logic of this world is the calculus of Probabilities,* which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man’s mind.

Harold Jeffreys (1931)

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.
## Interpreting Bayesian probabilities

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>
</table>
| Hot, cold        | Temperature, $T$ | Cold as ice $= 273K$  
Boiling hot $= 373K$ |
| uncertainty      | Probability, $P$ | Certainty $= 0, 1$  
$p = 1/36$: plausible as “snake’s eyes”  
$p = 1/1024$: plausible as 10 heads |
Interpreting PDFs

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:

![Bayesian Distribution](image)

Frequentist

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* in a sequence of trials. \( p(x) \) describes how the *values of* \( x \) would be distributed among infinitely many trials:

![Frequentist Distribution](image)
Agenda

1. Confidence intervals vs. credible intervals
2. Probability theory as generalized logic
3. Probability theory for data analysis: Three theorems
4. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty
5. Multilevel models for measurement error
6. Bayesian computation
The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information (including observed data) 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): $P(H_1, D_{\text{obs}}|I) = P(H_1|I)P(D_{\text{obs}}|H_1, I)$

$= P(D_{\text{obs}}|I)P(H_1|D_{\text{obs}}, I)$

‘NOT’: $P(H_1|I) = 1 - P(H_1|I)$
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 $D_{\text{obs}} = \text{Average likelihood for } H_i$
(a.k.a. marginal likelihood)

**Normalization**

For exclusive, exhaustive $H_i$,

$$\sum_i P(H_i | \cdots) = 1$$
Visualizing Bayesian Inference

Simplest case: Binary classification

- 2 hypotheses: \( \{ C, \overline{C} \} \)
- 2 possible data values: \( \{-, +\} \)

Concrete example: You test positive (+) for a medical condition. Do you have the condition (C) or not (\( \overline{C} \))?  

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
  - Test is 80% accurate if you have the condition:  
    \[ P(+|C, I) = 0.8 \] ("sensitivity")
  - Test is 95% accurate if you are healthy:  
    \[ P(-|\overline{C}, I) = 0.95 \] ("specificity," \( 1 - p(\text{false +}) \))

*Numbers roughly correspond to mammography screening for breast cancer in asymptomatic women*
Case diagram—probabilities

\[
P(C|+, I) = \frac{0.0008}{0.05075} \approx 0.016
\]

\[
P(H_1 \lor H_2|I)
\]

\[
P(H_i|I)
\]

\[
P(H_i, D_j|I) = P(H_i|I)P(D_j|H_i, I)
\]

\[
P(D_j|I) = \sum_i P(H_i, D_j|I)
\]
Case diagram—counts

Create a large ensemble of imaginary cases so ratios of counts approximate the probabilities.

Of the 508 cases with positive test results, only 8 have the condition. The prevalence is so low that when there is a positive result, it’s more likely to have been a mistake than accurate, even for a sensitive test.

\[ P(C|+, I) \approx \frac{8}{508} \approx 0.016 \]
Inference as manipulation of the joint distribution

Bayes’s theorem in terms of the \textit{joint distribution}:

\[
P(H_i|I) \times P(D_{\text{obs}}|H_i, I) = P(H_i, D_{\text{obs}}|I) = P(H_i|D_{\text{obs}}, I) \times P(D_{\text{obs}}|I)
\]
Joint
dist’n

Conditional dist’ns

Prior predictive

\[ \text{Joint} \]

\[ \text{Conditional dist’ns} \]

\[ \text{Prior predictive} \]

\[ D_{\text{obs}} \]

\[ \text{Posterior dist’n} \]
Prior Sampling dist'ns

Data

Hypotheses

Prior

Likelihood function

Joint

Conditional dist'ns

Posterior dist'n

Prior predictive

$D_{\text{obs}}$

$D_{\text{obs}}$

$D_{\text{obs}}$

$D_{\text{obs}}$
Continuous data, parameter spaces

Components of Bayes’s theorem for a problem with a 1-D parameter space (θ) and a 2-D sample space (y), with observed data \( y_d \), and modeling assumptions A.
Agenda

1. Confidence intervals vs. credible intervals
2. Probability theory as generalized logic
3. Probability theory for data analysis: Three theorems
4. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty
5. Multilevel models for measurement error
6. Bayesian computation
Models $M_i$ ($i = 1$ to $N$), each with parameters $\theta_i$, each imply a \textit{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 \textit{observed} data is the \textit{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 $\theta_i$ (parameter uncertainty).

\textit{Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript:} $D = D_{\text{obs}}$. 
Classes of Problems

Single-model inference

Premise = choice of single model (specific $i$)

*Parameter estimation*: What can we say about $\theta_i$ or $f(\theta_i)$?
*Prediction*: What can we say about future data $D'$?

Multi-model inference

Premise = $\{M_i\}$

*Model comparison/choice*: What can we say about $i$?
*Model averaging*:
  - *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; $\phi$ is common to all
    What can we say about $\phi$ w/o committing to one model?
  - *Prediction*: What can we say about future $D'$, accounting for model uncertainty?

Model checking

Premise = $M_1 \lor$ “all” alternatives
Is $M_1$ adequate? (predictive tests, calibration, robustness)
Parameter Estimation

Problem statement

\( I = \) Model \( M \) with parameters \( \theta \) (+ any add’l info)

\( H_i = \) statements about \( \theta \); 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 \( \theta \):

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

Posterior probability density

\[
p(\theta|D, M) = \frac{p(\theta|M) \mathcal{L}(\theta)}{\int d\theta \ p(\theta|M) \mathcal{L}(\theta)}
\]
**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** $\Delta$ 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 $\phi$, nuisance parameters $\eta$
  - *Marginal dist’n* for $\phi$: $p(\phi|D, M) = \int d\eta p(\phi, \eta|D, M)$
Estimating a Normal Mean

Problem specification

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

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

Likelihood

\[
\mathcal{L}(\mu) \equiv p(D|\mu, \sigma, M) = \prod_i \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(d_i-\mu)^2}{2\sigma^2}}; \quad \sigma = 1
\]

\[
\propto \exp \left( -\frac{N(\mu - \bar{d})^2}{2\sigma^2} \right)
\]

Likelihood function is a Gaussian function at \( \bar{d} \), width \( w = \sigma / \sqrt{N} \)
**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

$w_0 \to \infty$ is the “uninformative” flat prior limit; posterior remains normal and proper (normalizable)

**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

$$\tilde{\mu} = \bar{d} + B \cdot (\mu_0 - \bar{d})$$

$$\tilde{w} = w \cdot \sqrt{1 - B}$$

“*Principle of stable estimation*” — The prior affects estimates only when data are not informative relative to prior
Conjugate normal examples:

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

- Binomial example
  - Bernoulli trials, binomial & negative binomial dist’ns
  - Beta-binomial conjugate model
  - Likelihood principles

- Normal example
  - Analytical details for normal example
  - Sufficiency; sample mean and variance as sufficient statistics
  - Connection to least-squares curve fitting
  - Handling $\sigma$ uncertainty by marginalizing over $\sigma$; Student’s $t$ distribution
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

To summarize implications for $s$, accounting for $b$ uncertainty, marginalize:

$$
p(s|D, M) = \int db \; p(s, b|D, M)
$$

$$
\propto p(s|M) \int db \; p(b|s, M) \mathcal{L}(s, b)
$$

$$
= p(s|M) \mathcal{L}_m(s)
$$

with $\mathcal{L}_m(s)$ the marginal likelihood function for $s$:

$$
\mathcal{L}_m(s) \equiv \int db \; p(b|s) \mathcal{L}(s, b)
$$
Marginalization vs. Profiling

For insight: Suppose the prior is broad compared to the likelihood → for a fixed $s$, we can accurately estimate $b$ with max likelihood $\hat{b}_s$, with small uncertainty $\delta b_s$.

$$\mathcal{L}_m(s) \equiv \int db \ p(b|s) \mathcal{L}(s, b)$$

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

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

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}, \quad \sigma^2_s = \sigma^2_r + \sigma^2_b$

Background subtraction is a special case of background marginalization.
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$

$\delta \hat{b}_s$ is const. vs. $s$

$\Rightarrow \mathcal{L}_m \propto \mathcal{L}_p$
Flared/skewed/banana-shaped: $\mathcal{L}_m$ and $\mathcal{L}_p$ differ

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

In “measurement error problems” the difference can be dramatic
The On/Off Problem for Poisson counting data

Basic problem

- Look off-source; unknown background rate $b$
  Count $N_{\text{off}}$ photons in interval $T_{\text{off}}$

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

- Infer $s$

Conventional solution

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

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)
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

The likelihood function is a product of separate Poisson distributions for the off-source and on-source data:

$$
\mathcal{L}(s, b) = \left( \frac{b \cdot T_{\text{off}}}{N_{\text{off}}} \right)^{N_{\text{off}}} \times \left( \frac{(s + b) \cdot T_{\text{on}}}{N_{\text{on}}} \right)^{N_{\text{on}}} \times e^{-b \cdot T_{\text{off}}} \times e^{-(s+b) \cdot T_{\text{on}}}
$$

Adopting flat priors for \((s, b)\), the joint posterior is

$$
p(s, b|N_{\text{on}}, N_{\text{off}}, I) \propto (s + b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-s \cdot T_{\text{on}}} e^{-b \cdot (T_{\text{on}} + T_{\text{off}})}
$$

If \(b = 0\), the (normalized) posterior distribution is a gamma distribution,

$$
p(s, b = 0|N_{\text{on}}, N_{\text{off}}, I) = \frac{T_{\text{on}}(s \cdot T_{\text{on}})^{N_{\text{on}}}}{N_{\text{on}}!} e^{-s \cdot T_{\text{on}}}
$$
Now marginalize over $b$;

$$p(s|N_{\text{on}}, N_{\text{off}}, l) = \int db \ p(s, b | N_{\text{on}}, l_{\text{all}})$$

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

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

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

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

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

\[ T_{\text{on}} = 1 \]

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

\[ N_{\text{on}} = 6 \]

\[ N_{\text{on}} = 9 \]

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

\[ T_{\text{on}} = 1 \]

\[ T_{\text{off}} = 1, \quad N_{\text{off}} = 9 \]
\[ T_{\text{off}} = 10, \quad N_{\text{off}} = 90 \]
Supplement:

- Analytical details for Poisson dist’n inference
- Gamma-Poisson conjugate model
- Alternative (equivalent) solutions to the on/off problem
- Multibin case
Many Roles for Marginalization

Eliminate nuisance parameters

\[ p(\phi|D, M) = \int d\eta \ p(\phi, \eta|D, M) \]

Propagate uncertainty

Model has parameters \( \theta \); what can we infer about \( F = f(\theta) \)?

\[
\begin{align*}
    p(F|D, M) &= \int d\theta \ p(F, \theta|D, M) = \int d\theta \ p(\theta|D, M) p(F|\theta, M) \\
    &= \int d\theta \ p(\theta|D, M) \delta[F - f(\theta)] \quad [\text{single-valued case}]
\end{align*}
\]

Prediction

Given a model with parameters \( \theta \) and present data \( D \), predict future data \( D' \) (e.g., for experimental design):

\[
\begin{align*}
    p(D'|D, M) &= \int d\theta \ p(D', \theta|D, M) = \int d\theta \ p(\theta|D, M) p(D'|\theta, M)
\end{align*}
\]

Model comparison...
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**

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

\[ \propto p(M_i|I)L(M_i) \]

\[
L(M_i) = p(D|M_i) = \int d\theta_i p(\theta_i|M_i)p(D|\theta_i, M_i).
\]

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:

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

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

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

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods.
An Automatic Ockham’s Razor

Consider *nested models*:

- Simpler model $M_1$ with parameters $\theta_1$
- “Larger” rival $M_2$ with parameters $\theta_2 = (\theta_1, \eta)$

$\Rightarrow \mathcal{L}(\hat{\theta}_2) \geq \mathcal{L}(\hat{\theta}_1)$

But what about $p(D|M_i) = \int d\theta_i \ p(\theta_i|M) \ \mathcal{L}(\theta_i)$?

*Prior predictive distributions*

Normalization implies *there must be data that favor $M_1$*:
The Ockham Factor

\[ p(D|M_i) = \int d\theta_i \ p(\theta_i|M) \ \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M) \mathcal{L}(\hat{\theta}_i) \delta\theta_i \approx \mathcal{L}(\hat{\theta}_i) \frac{\Delta\theta_i}{\Delta\theta_i} = \text{Maximum Likelihood} \times \text{Ockham Factor} \]

Models with more parameters often make the data more probable — for the best fit

Ockham factor penalizes models for “wasted” volume of parameter space

Quantifies intuition that models shouldn’t require fine-tuning
Issues In Coincidence Assessment

• **Choice of statistic** — How to measure “close”? 
  *Nearest neighbor distance, correlation functions, Mahalanobis distance, likelihood ratio, Bayes factor*

• **Proximity criterion** — How close is “close enough”? 
  *p-value/significance level, power, odds/Bayes factor*

• **Tuning/fitting issues:**
  ▶ Tuning of statistics/parameters — *Sequential analysis; marginalization* 
  ▶ Number of candidates — *Multiple testing; model averaging* 
  ▶ Choice of counterpart catalogs — *Inherently domain-specific?*

• **Computation with large catalogs**

  *IVOA Open SkyQuery addresses these using Bayesian model comparison*

Axisymmetric uncertainties: Fisher distributions

Let \( \hat{n}_i \) be the best-fit direction (2 location parameters). For azimuthally-symmetric uncertainties, use:

\[
\mathcal{L}_i(n) = \frac{\kappa_i}{4\pi \sinh \kappa_i} e^{\kappa_i n \cdot \hat{n}_i}
\]

\( \kappa_i = \) concentration parameter. For small uncertainties,

\[
\kappa_i \approx \frac{C}{\sigma_i^2}, \quad C \approx 2.3
\]

If \( n \) is near \( \hat{n}_i \) (separation angle \( \theta \))

\[
\mathcal{L}_i(n) \sim \exp \left[ -\frac{C \theta^2}{2\sigma_i^2} \right]
\]

Other Gaussian-like properties: sufficient statistics, maximum entropy

Non-Axisymmetric: Kent (Fisher-Bingham), Fisher mixture...
Bayesian Coincidence Assessment

Not associated

\[ n_1 \xrightarrow{} D_1 \]
\[ n_2 \xrightarrow{} D_2 \]

Associated

\[ n \]
\[ D_1 \]
\[ D_2 \]

\[ \ell_1(n_1) \quad \ell_2(n_2) \]

\[ p(d_1, d_2 | H_0) = \int dn_1 p(n_1 | H_0) \ell_1(n_1) \]
\[ \times \int dn_2 \cdots \]

\[ p(d_1, d_2 | H_1) = \int dn p(n | H_1) \ell_1(n) \ell_2(n) \]
Doublet Bayes factor behavior vs. nearest-neighbor \( p \)-value

Separation angle (°)

Bayes factor

\( p \)-value = 0.05

\( p \)-value = 0

\( \sigma = 10° \)

\( \sigma = 25° \)
Challenge: Large hypothesis spaces

For $N = 2$ events, there was a single coincidence hypothesis, $H_1$

For $N = 3$ events:

- Three doublets: $1 + 2$, $1 + 3$, or $2 + 3$
- One triplet

The number of alternatives (partitions, $\varpi$) grows combinatorially!

- **Model building**: Assign sensible priors to partitions
- **Computation**: Find & sum over important partitions
Theme: Parameter Space Volume

Bayesian calculations *sum/integrate over parameter/hypothesis space*

(Frequentist calculations average over *sample space* & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space.
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters.
- Model likelihoods have Ockham 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” → account for size of parameter space

Physical analogy

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

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

Maximum likelihood focuses on the “hottest” parameters. Bayes focuses on the parameters 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.
Supplement:

- Assigning priors
- Rule-based non-informative priors: Jeffreys, reference
Recap of Key Ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

*Bayes’s theorem*

\[ p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis}) \]

Data *change* the support for a hypothesis \( \propto \) ability of hypothesis to *predict* the data

*Law of total probability*

\[ p(\text{Hypotheses} \mid \text{Data}) = \sum p(\text{Hypothesis} \mid \text{Data}) \]

The support for a *compound/composite* hypothesis must account for all the ways it could be true
Agenda

1. Confidence intervals vs. credible intervals
2. Probability theory as generalized logic
3. Probability theory for data analysis: Three theorems
4. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty
5. Multilevel models for measurement error
6. Bayesian computation
Complications With Survey Data

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

- **“Scatter” effects** (measurement error, etc.) — *insidious*
  Typically ignored (average out???)
Many Guises of Measurement Error

Auger data above GZK cutoff (PAO 2007; Soiaporn + 2013)

QSO hardness vs. luminosity (Kelly 2007, 2012)
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

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

$$
p(\theta|\{x_i\}) \propto p(\theta)\mathcal{L}(\theta) = p(\theta, \{x_i\})
$$

(A binomial point process)
**Graphical representation**

- Nodes/vertices = uncertain quantities (gray → known)
- Edges specify conditional dependence
- Absence of an edge denotes *conditional independence*

Graph specifies the form of the *joint distribution*:

\[ p(\theta, \{x_i\}) = p(\theta) p(\{x_i\}|\theta) = p(\theta) \prod_i f(x_i|\theta) \]

Posterior from BT:

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

$\{x_i\}$ are now uncertain (latent) parameters. We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$:

$$p(\theta, \{x_i\}, \{D_i\}) = p(\theta) p(\{x_i\}|\theta) p(\{D_i\}|\{x_i\})$$

$$= p(\theta) \prod_i f(x_i|\theta) \ell_i(x_i)$$

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

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

Key point: *Maximizing over $x_i$ and integrating over $x_i$ can give very different results!*
To estimate $x_1$:

\[
p(x_1|\{x_2, \ldots\}) = \int d\theta \ p(\theta) f(x_1|\theta) \ell_1(x_1) \times \prod_{i=2}^{N} \int dx_i \ f(x_i|\theta) \ell_i(x_i)
\]

\[
= \ell_1(x_1) \int d\theta \ p(\theta) f(x_1|\theta) \mathcal{L}_{m,\hat{1}}(\theta)
\]

\[
\approx \ell_1(x_1) f(x_1|\hat{\theta})
\]

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

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

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

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

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

\[
p(\theta, \{x_i\}, \{D_i\}) = p(\theta) \ p(\{x_i\}|\theta) \ p(\{D_i\}|\{x_i\}) \\
= p(\theta) \prod_i f(x_i|\theta) \ p(D_i|x_i) = p(\theta) \prod_i f(x_i|\theta) \ell_i(x_i)
\]
Bayesian MLMs in Astronomy

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

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

Directional & spatio-temporal coincidences:

► GRB repetition (Luo\(^+\) 1996; Graziani\(^+\) 1996)
► GRB host ID (Band 1998; Graziani\(^+\) 1999)
► VO cross-matching (Budavári & Szalay 2008)
Time series:

▶ SN 1987A neutrinos, uncertain energy vs. time (Loredo & Lamb 2002)

▶ Multivariate “Bayesian Blocks” (Dobigeon, Tourneret & Scargle 2007)

▶ SN Ia multicolor light curve modeling (Mandel+ 2009, 2011)

Linear & nonlinear regression with measurement error:

▶ QSO hardness vs. luminosity (Kelly 2007, 2012)

▶ Dust SEDs (Kelly+ 2012)

More information:

CASSt 2014 Supplement Session

Overview of MLMs in astronomy: arXiv:1208.3036

In progress: GPU software (Szalai-Gindl, Budavari, Kelly, TL)
Agenda

1. Confidence intervals vs. credible intervals
2. Probability theory as generalized logic
3. Probability theory for data analysis: Three theorems
4. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty
5. Multilevel models for measurement error
6. Bayesian computation
Parameter space integrals

For model with $m$ parameters, we need to evaluate integrals like:

$$
\int d^m \theta \ g(\theta) \pi(\theta) \mathcal{L}(\theta) = \int d^m \theta \ g(\theta) \ q(\theta)
$$

- $g(\theta) = 1 \rightarrow Z = p(D|M)$ (norm. const., model likelihood)
- $g(\theta) = \theta \rightarrow$ posterior mean for $\theta$
- $g(\theta) =$ ‘box’ $\rightarrow$ probability $\theta \in$ credible region
- $g(\theta) = 1$, integrate over subspace $\rightarrow$ marginal posterior
- $g(\theta) = \delta[\psi - \psi(\theta)] \rightarrow$ propagate uncertainty to $\psi(\theta)$

Except for optimization, Bayesian computation amounts to computing the expectation of some function $g(\theta)$ with respect to the posterior dist’n for $\theta$

Contrast with frequentist computation, which integrates over sample space, e.g., via Monte Carlo simulation of data
Bayesian Computation Menu

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

Modest-dimensional models ($m \lesssim 10$ to 20)
- Quadrature, cubature, adaptive cubature
- IID Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($m \gtrsim 5$): Non-IID Monte Carlo
- Posterior sampling — create RNG that samples posterior
  - Markov Chain Monte Carlo (MCMC) is the most general framework
- Sequential Monte Carlo (SMC)
- Approximate Bayesian computation (ABC)
- ...

More in presentations by Murali Haran, Eric Ford!
See SCMA 5 Bayesian Computation tutorial notes, and notes from CASt 2014 Supplement Sessions, for more on computation!

See online resource list for an annotated list of Bayesian books and software.