Introduction to Bayesian Inference
Lecture 1: Fundamentals

Tom Loredo
Dept. of Astronomy, Cornell University
http://www.astro.cornell.edu/staff/loredo/bayes/
Lectures http://inference.astro.cornell.edu/INPE09/

CASt Summer School — 11 June 2010

Lecture 1: Fundamentals

1. The big picture
2. A flavor of Bayes: χ^2 confidence/credible regions
3. Foundations: Logic & probability theory
4. Probability theory for data analysis: Three theorems
5. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

Bayesian Fundamentals

1. The big picture
2. A flavor of Bayes: χ^2 confidence/credible regions
3. Foundations: Logic & probability theory
4. Probability theory for data analysis: Three theorems
5. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

Scientific Method

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

Scientists *argue!*

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

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

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

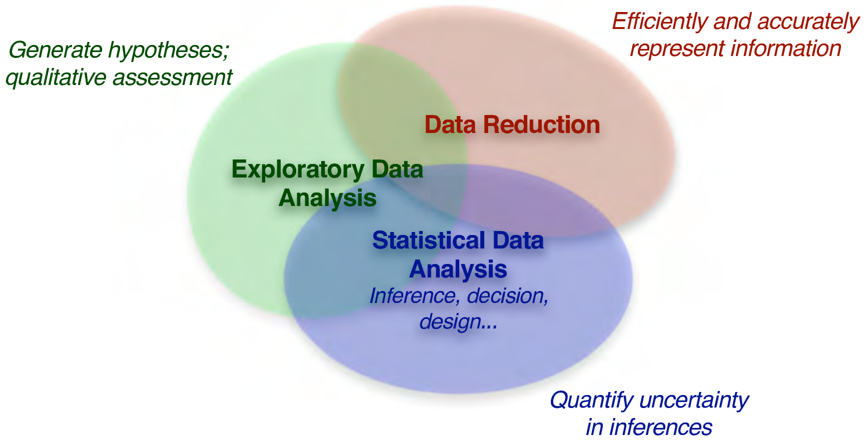
We gather data so they may speak for or against existing hypotheses, and guide the formation of new hypotheses.

A key role of data in science is to be among the premises in scientific arguments.

Data Analysis

Building & Appraising Arguments Using Data

Modes of Data Analysis



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, χ^2 testing, ANOVA, survival analysis . . .)
- Foundation: Use probability theory to quantify the strength of arguments (i.e., a more abstract view than restricting PT to describe variability in repeated "random" experiments)
- Focuses on *deriving consequences of modeling assumptions* rather than *devising and calibrating procedures*

Frequentist vs. Bayesian Statements

"I find conclusion C based on data $D_{\rm obs}$ . . . "

Frequentist assessment

"It was found with a procedure that's right 95% of the time over the set $\{D_{\rm hyp}\}$ that includes $D_{\rm obs}$."
Probabilities are properties of *procedures*, not of particular results.

Bayesian assessment

"The strength of the chain of reasoning from $D_{\rm obs}$ to C is 0.95, on a scale where 1= certainty."
Probabilities are properties of *specific results*.
Long-run performance must be separately evaluated (and is typically good by frequentist criteria).

Bayesian Fundamentals

1. The big picture
2. A flavor of Bayes: χ^2 confidence/credible regions
3. Foundations: Logic & probability theory
4. Probability theory for data analysis: Three theorems
5. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

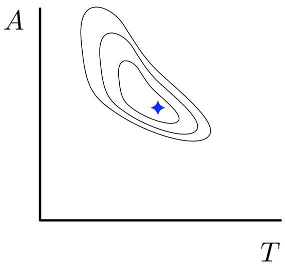
Estimating Parameters Via χ^2

Collect data $D_{\text{obs}} = \{d_i, \sigma_i\}$, fit with 2-parameter model via χ^2:

$$\chi^2(A, T) = \sum_{i=1}^{N} \frac{[d_i - f_i(A, T)]^2}{\sigma_i^2}$$

Two classes of variables

- Data (samples) d_i — Known, define N-D *sample space*
- Parameters $\theta = (A, T)$ — Unknown, define 2-D *parameter space*

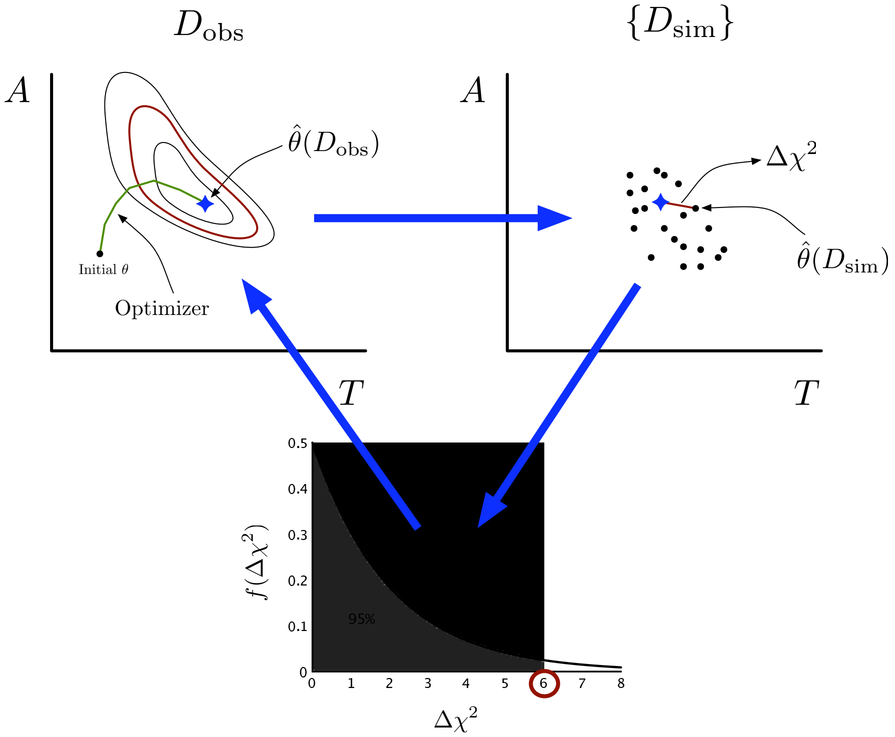


"Best fit" $\hat{\theta} = \arg\min_\theta \chi^2(\theta)$

Report uncertainties via χ^2 contours, but how do we quantify uncertainty vs. contour level?

Page 266

Frequentist: Parametric Bootstrap



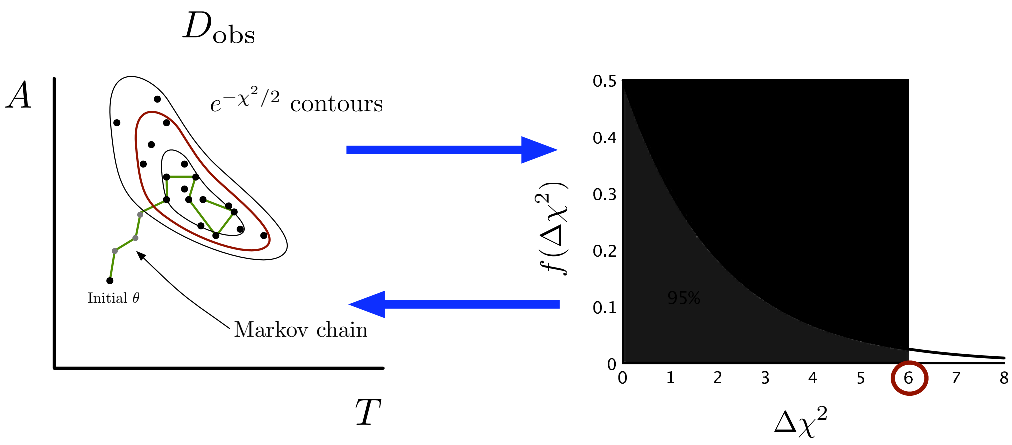
Parametric Bootstrap Algorithm

Monte Carlo algorithm for finding approximate confidence regions:

1. Pretend true params $\theta^* = \hat{\theta}(D_{\rm obs})$ ("plug-in approx'n")
2. Repeat N times:
    1. Simulate a dataset from $p(D_{\rm sim}|\theta^*)$ $\rightarrow \chi^2_{D_{\rm sim}}(\theta)$
    2. Find min χ^2 estimate $\hat{\theta}(D_{\rm sim})$
    3. Calculate $\Delta\chi^2 = \chi^2(\theta^*) - \chi^2(\hat{\theta}_{D_{\rm sim}})$
4. Histogram the $\Delta\chi^2$ values to find coverage vs. $\Delta\chi^2$ (fraction of sim'ns with smaller $\Delta\chi^2$)

Result is approximate even for $N \rightarrow \infty$ because $\theta^* \neq \hat{\theta}(D_{\rm obs})$.

Bayesian: Posterior Sampling Via MCMC



Posterior Sampling Algorithm

Monte Carlo algorithm for finding credible regions:

1. Create a RNG that can sample θ from $p(\theta|D_{\rm obs}) \propto e^{-\chi^2(\theta)/2}$
2. Draw N samples; record θ_i and $q_i = \chi^2(\theta_i)$
3. Sort the samples by the q_i values
4. A credible region of probability P is the θ region spanned by the $100P\%$ of samples with highest q_i

Note that no dataset other than $D_{\rm obs}$ is ever considered.

The only approximation is the use of Monte Carlo.

These very different procedures produce the same regions for linear models with Gaussian error distributions.

These procedures produce *different* regions in general.

Bayesian Fundamentals

1. The big picture
2. A flavor of Bayes: χ^2 confidence/credible regions
3. Foundations: Logic & probability theory
4. Probability theory for data analysis: Three theorems
5. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

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 : $\Omega_{\rm tot} \in [0.9, 1.1]$

  B : Ω_Λ is not 0

  $\overline{B}$: "not B," i.e., $\Omega_\Lambda = 0$

- *Connectives:*

  $A \wedge B$: A and B are *both* true

  $A \vee B$: 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:

$H|\mathcal{P}$: Premises $\mathcal{P}$ 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 → 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 address *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.

Factual Correctness

Passing the buck

> 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*

An open issue

> 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

Premises define a fixed *context* in which arguments may be assessed.

Premises are considered "given"—if only for the sake of the 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.

Deductive Logic

Assess arguments by decomposing them into parts via connectives, and assessing the parts

Validity of $A \wedge B|\mathcal{P}$

	$A\|\mathcal{P}$	$\overline{A}\|\mathcal{P}$
$B\|\mathcal{P}$	valid	invalid
$\overline{B}\|\mathcal{P}$	invalid	invalid

Validity of $A \vee B|\mathcal{P}$

	$A\|\mathcal{P}$	$\overline{A}\|\mathcal{P}$
$B\|\mathcal{P}$	valid	valid
$\overline{B}\|\mathcal{P}$	valid	invalid

Representing Deduction With $\{0, 1\}$ Algebra

$V(H|\mathcal{P}) \equiv$ Validity of argument $H|\mathcal{P}$:

$$\begin{aligned} V &= 0 \longrightarrow \text{Argument is } \textit{invalid} \\ &= 1 \longrightarrow \text{Argument is } \textit{valid} \end{aligned}$$

Then deduction can be reduced to integer multiplication and addition over $\{0, 1\}$ (as in a computer):

$$\begin{aligned} V(A \wedge B|\mathcal{P}) &= V(A|\mathcal{P})\, V(B|\mathcal{P}) \\ V(A \vee B|\mathcal{P}) &= V(A|\mathcal{P}) + V(B|\mathcal{P}) - V(A \wedge B|\mathcal{P}) \\ V(\overline{A}|\mathcal{P}) &= 1 - V(A|\mathcal{P}) \end{aligned}$$

Representing Induction With $[0, 1]$ Algebra

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

$$\begin{aligned} P &= 0 \longrightarrow \text{Argument is } \textit{invalid}\text{; premises imply } \overline{H} \\ &= 1 \longrightarrow \text{Argument is } \textit{valid} \\ &\in (0, 1) \longrightarrow \text{Degree of deducibility} \end{aligned}$$

Mathematical model for induction

'AND' (product rule): $$\begin{aligned} P(A \wedge B|\mathcal{P}) &= P(A|\mathcal{P})\, P(B|A \wedge \mathcal{P}) \\ &= P(B|\mathcal{P})\, P(A|B \wedge \mathcal{P}) \end{aligned}$$

'OR' (sum rule): $$\begin{aligned} P(A \vee B|\mathcal{P}) &= P(A|\mathcal{P}) + P(B|\mathcal{P}) \\ &\quad - P(A \wedge B|\mathcal{P}) \end{aligned}$$

'NOT': $$P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$$

The Product Rule

We simply promoted the V algebra to real numbers; the only thing changed is part of the product rule:

$$\begin{aligned} V(A \wedge B|\mathcal{P}) &= V(A|\mathcal{P})\, V(B|\mathcal{P}) \\ P(A \wedge B|\mathcal{P}) &= P(A|\mathcal{P})\, P(B|A, \mathcal{P}) \end{aligned}$$

Suppose A implies B (i.e., $B|A, \mathcal{P}$ is valid). Then we don't expect $P(A \wedge B|\mathcal{P})$ to differ from $P(A|\mathcal{P})$.

In particular, $P(A \wedge A|\mathcal{P})$ must equal $P(A|\mathcal{P})$!

Such qualitative reasoning satisfied early probabilists that the sum and product rules were worth considering as axioms for a theory of quantified induction.

Firm Foundations

Today many different lines of argument *derive* induction-as-probability from various simple and appealing requirements:

- Consistency with logic + internal consistency (Cox; Jaynes)
- "Coherence"/optimal betting (Ramsey; DeFinetti; Wald; Savage)
- Algorithmic information theory (Rissanen; Wallace & Freeman)
- Optimal information processing (Zellner)
- Avoiding problems with frequentist methods:
  - Avoiding recognizable subsets (Cornfield)
  - Avoiding stopping rule problems → likelihood principle (Birnbaum; Berger & Wolpert)

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*

More On Interpretation

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."

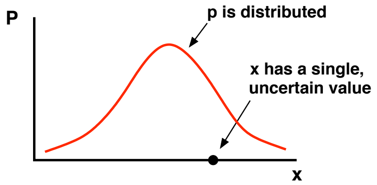
A Thermal Analogy

Intuitive notion	*Quantification*	*Calibration*
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

A Bit More On Interpretation

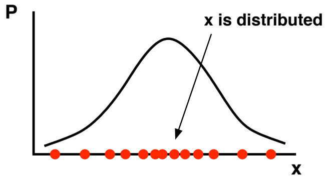
Bayesian

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:



Frequentist

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



Bayesian Fundamentals

1. The big picture
2. A flavor of Bayes: χ^2 confidence/credible regions
3. Foundations: Logic & probability theory
4. Probability theory for data analysis: Three theorems
5. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

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: $H_i|D_{\rm obs}, I$. ($I =$ "background information," everything deemed relevant besides the observed data)

$P(H_i|D_{\rm obs}, I)$ measures the degree to which $(D_{\rm obs}, I)$ 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 $P(H_i|D_{\rm obs}, I)$ 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:

$$\text{`OR' (sum rule):}\quad P(H_1 \vee H_2|I) = P(H_1|I) + P(H_2|I) - P(H_1, H_2|I)$$

$$\begin{aligned}\text{`AND' (product rule):}\quad P(H_1, D|I) &= P(H_1|I)\, P(D|H_1, I)\\ &= P(D|I)\, P(H_1|D, I)\end{aligned}$$

$$\text{`NOT':}\quad P(\overline{H_1}|I) = 1 - P(H_1|I)$$

Three Important Theorems

Bayes's Theorem (BT)

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

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

Solve for the *posterior probability*:

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

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

posterior ∝ *prior* × *likelihood*

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

Law of Total Probability (LTP)

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

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

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

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

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

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

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

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

prior predictive for D_{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”).

Visualizing Bayesian Inference

Simplest case: Binary classification

- 2 hypotheses: $\{H, C\}$
- 2 possible data values: $\{-, +\}$

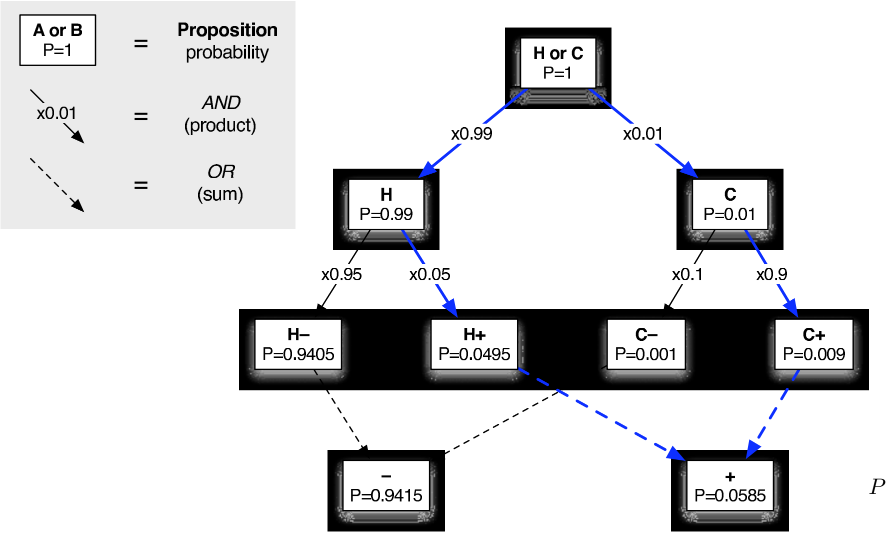
Concrete example: You test positive $(+)$ for a medical condition. Do you have the condition (C) or not $(H$, "healthy")?

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
  - Test is 90% accurate if you have the condition: $P(+|C, I) = 0.9$ ("sensitivity")
  - Test is 95% accurate if you are healthy: $P(-|H, I) = 0.95$ ("specificity")

Numbers roughly correspond to breast cancer in asymptomatic women aged 40–50, and mammography screening

[Gigerenzer, *Calculated Risks* (2002)]

Probability Lattice



$P(H_1 \vee H_2|I)$

$P(H_i|I)$

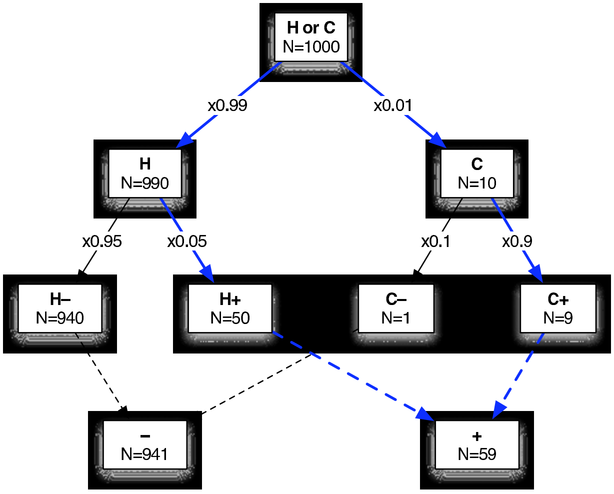
$P(H_i, D|I) = P(H_i|I)P(D|H_i, I)$

$P(D|I) = \sum_i P(H_i, D|I)$

$$P(C|+, I) = \frac{0.009}{0.0585} \approx 0.15$$

Count Lattice

Integers are easier than reals!
Create a large ensemble of cases so ratios of counts approximate the probabilities.



H or C
N=1000

x0.99

x0.01

H
N=990

C
N=10

x0.95

x0.05

x0.1

x0.9

H–
N=940

H+
N=50

C–
N=1

C+
N=9

–
N=941

+
N=59

$$P(C|+,I) = \frac{9}{59} \approx 0.15$$

Of the 59 cases with positive test results, only 9 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.

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: To adjust your appraisal when new evidence becomes available, add the evidence to your initial premises.
- LTP: If the premises allow multiple arguments for a hypothesis, its appraisal must account for all of them.

Bayesian Fundamentals

1. The big picture
2. A flavor of Bayes: χ^2 confidence/credible regions
3. Foundations: Logic & probability theory
4. Probability theory for data analysis: Three theorems
5. Inference with parametric models
   - Parameter Estimation
   - Model Uncertainty

Inference With Parametric Models

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

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

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

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

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

Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript: $D = D_{obs}$.

Page 282

Three Classes of Problems

Parameter Estimation

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

Model Assessment

- Model comparison: Premise = $\{M_i\}$
  $\rightarrow$ What can we say about i?
- Model adequacy/GoF: Premise = $M_1 \vee$ "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 ϕ w/o committing to one model?
(Examples: systematic error, prediction)

Parameter Estimation

Problem statement

I = Model M with parameters θ (+ any add'l info)

H_i = statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"

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

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

Posterior probability density

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

Summaries of posterior

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

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

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

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

$$\mathcal{L}_m(s) \approx p(\hat{b}_s|s)\;\mathcal{L}(s,\overset{\text{best } b \text{ given } s}{\hat{b}_s})\;\underset{b \text{ uncertainty given } s}{\delta b_s}$$

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*.

Marginalization vs. Profiling

Marginal distribution for signal s, eliminating backgrond b:

$$p(s|D,M) \propto p(s|M)\mathcal{L}_m(s)$$

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

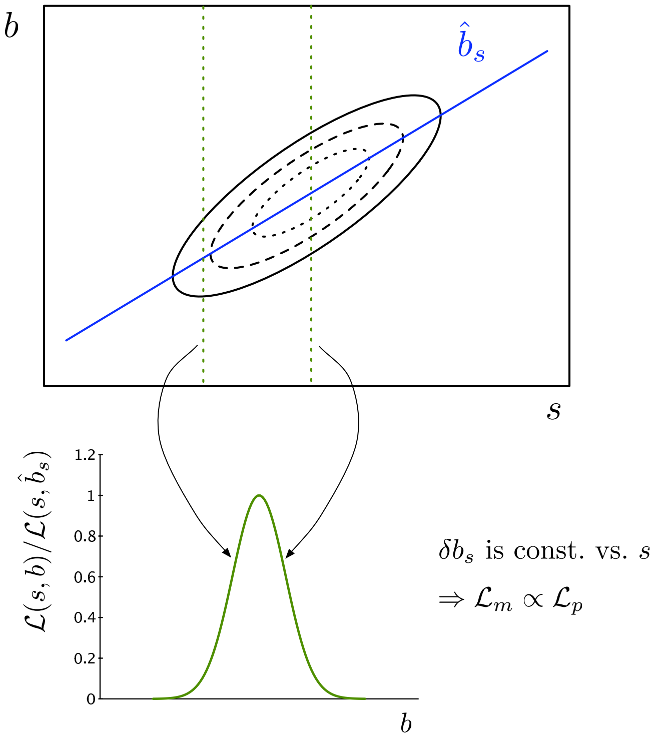
$$\mathcal{L}_m(s) \equiv \int db\; p(b|s)\,\mathcal{L}(s,b)$$

For insight: Suppose for a fixed s, we can accurately estimate b with max likelihood $\hat{b}_s$, with small uncertainty δb_s.

$$\begin{aligned} \mathcal{L}_m(s) &\equiv \int db\; p(b|s)\,\mathcal{L}(s,b) \\ &\approx p(\hat{b}_s|s)\;\mathcal{L}(s,\overset{\text{best } b \text{ given } s}{\hat{b}_s})\;\underset{b \text{ uncertainty given } s}{\delta b_s} \end{aligned}$$

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

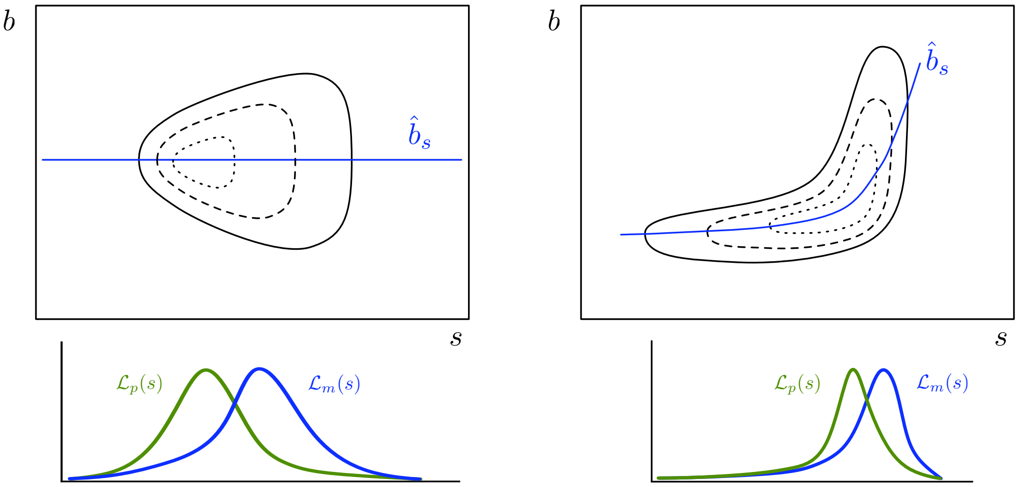
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



δb_s is const. vs. s

$\Rightarrow \mathcal{L}_m \propto \mathcal{L}_p$

Flared/skewed/bannana-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* (future lecture!) the difference can be dramatic.

Many Roles for Marginalization

Eliminate nuisance parameters

$$p(\phi|D, M) = \int d\eta\; p(\phi, \eta|D, M)$$

Propagate uncertainty

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

$$\begin{aligned} p(F|D, M) &= \int d\theta\; p(F, \theta|D, M) = \int d\theta\; p(\theta|D, M)\, p(F|\theta, M) \\ &= \int d\theta\; p(\theta|D, M)\, \delta[F - f(\theta)] \qquad \text{[\textit{single-valued case}]} \end{aligned}$$

Prediction

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

$$p(D'|D, M) = \int d\theta\; p(D', \theta|D, M) = \int d\theta\; p(\theta|D, M)\, p(D'|\theta, M)$$

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

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

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

Likelihood for model = Average likelihood for its parameters

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

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

Odds and Bayes factors

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

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

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

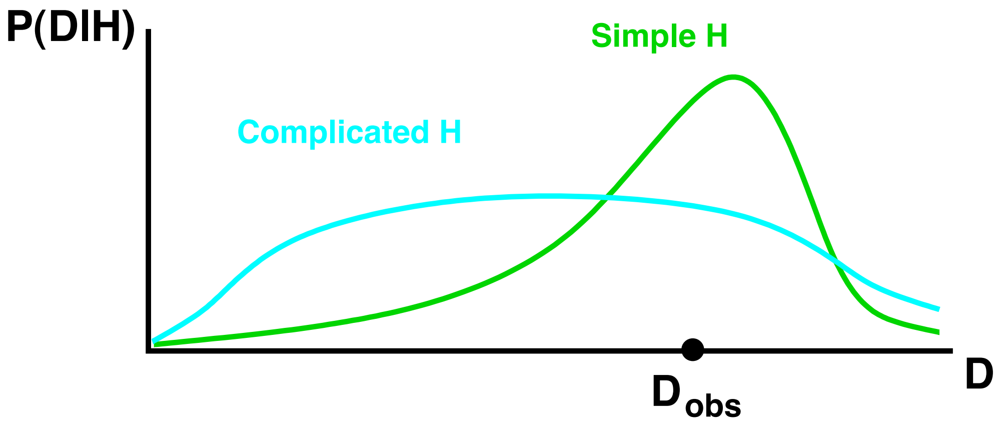
$$B_{ij} \equiv \frac{p(D|M_i,I)}{p(D|M_j,I)}$$

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

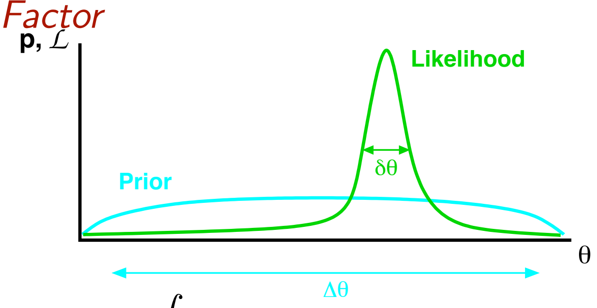
An Automatic Occam's Razor

Predictive probabilities can favor simpler models

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



The Occam Factor



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

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

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

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, ϕ
Each has different set of nuisance parameters η_i (or different prior info about them)
H_i = statements about ϕ

Model averaging

Calculate posterior PDF for ϕ:

$$
\begin{aligned}
p(\phi|D,I) &= \sum_i p(M_i|D,I)\, p(\phi|D,M_i) \\
&\propto \sum_i \mathcal{L}(M_i) \int d\eta_i \, p(\phi,\eta_i|D,M_i)
\end{aligned}
$$

The model choice is a (discrete) nuisance parameter here.

Theme: Parameter Space Volume

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

$$\text{Heat:} \quad Q = \int dV\, c_v(\mathbf{r}) T(\mathbf{r})$$

$$\text{Probability:} \quad P \propto \int d\theta\, p(\theta|I)\mathcal{L}(\theta)$$

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
    - Do not integrate/average over hypothetical data