Bayesian Inference in Astronomy & Astrophysics

Lecture 1: Fundamentals

Tom Loredo

Dept. of Astronomy, Cornell University

Today’s Lectures

- Fundamentals
- Inference with counting& point process models
- Probability and frequency

Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Scientific Method

Scientists *argue!*

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

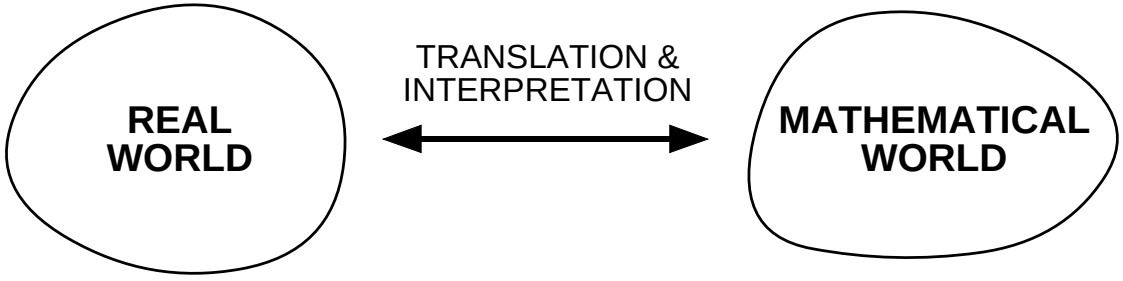
A key goal: Explain or predict *quantitative measurements*

Framework: Mathematical modeling

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

Mathematical Models

A model (in physics) is a representation of structure in a physical system and/or its properties. (Hestenes)



A model is a surrogate

The model is not the modeled system! It “stands in” for a particular purpose, and is subject to revision.

A model is an idealization

A model is a *caricature* of the system being modeled (Kac). It focuses on a subset of system properties of interest.

A model is an abstraction

Models identify common features of different things so that general ideas can be created and applied to different situations.

We seek a mathematical model for quantifying uncertainty—it will share these characteristics with physical models.

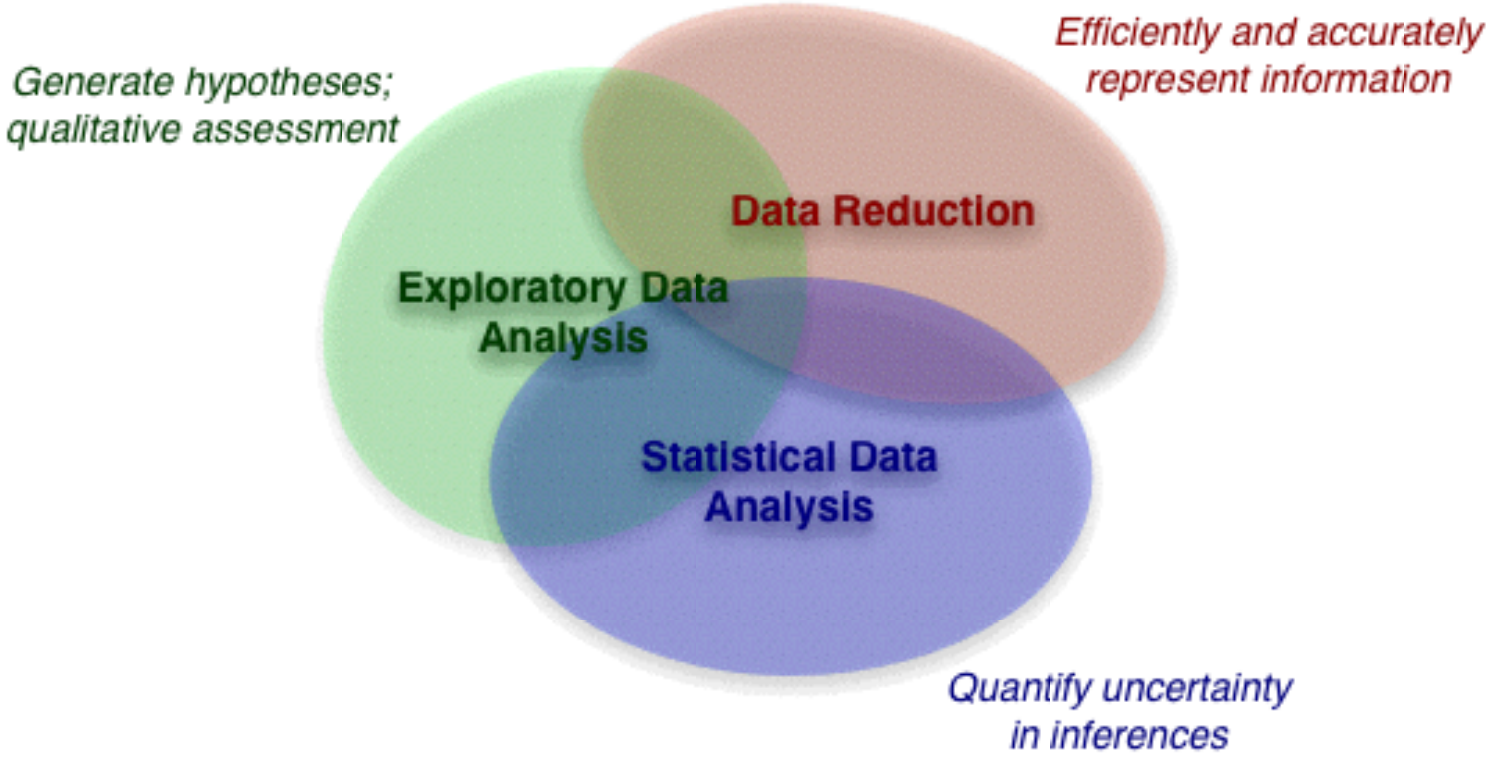
Asides

Theories are frameworks guiding model construction (laws, principals).

Physics as modeling is a leading school of thought in physics education research; e.g., `http://modeling.la.asu.edu/`

Data Analysis

Building & Appraising Arguments Using Data



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

Bayesian Statistical Inference

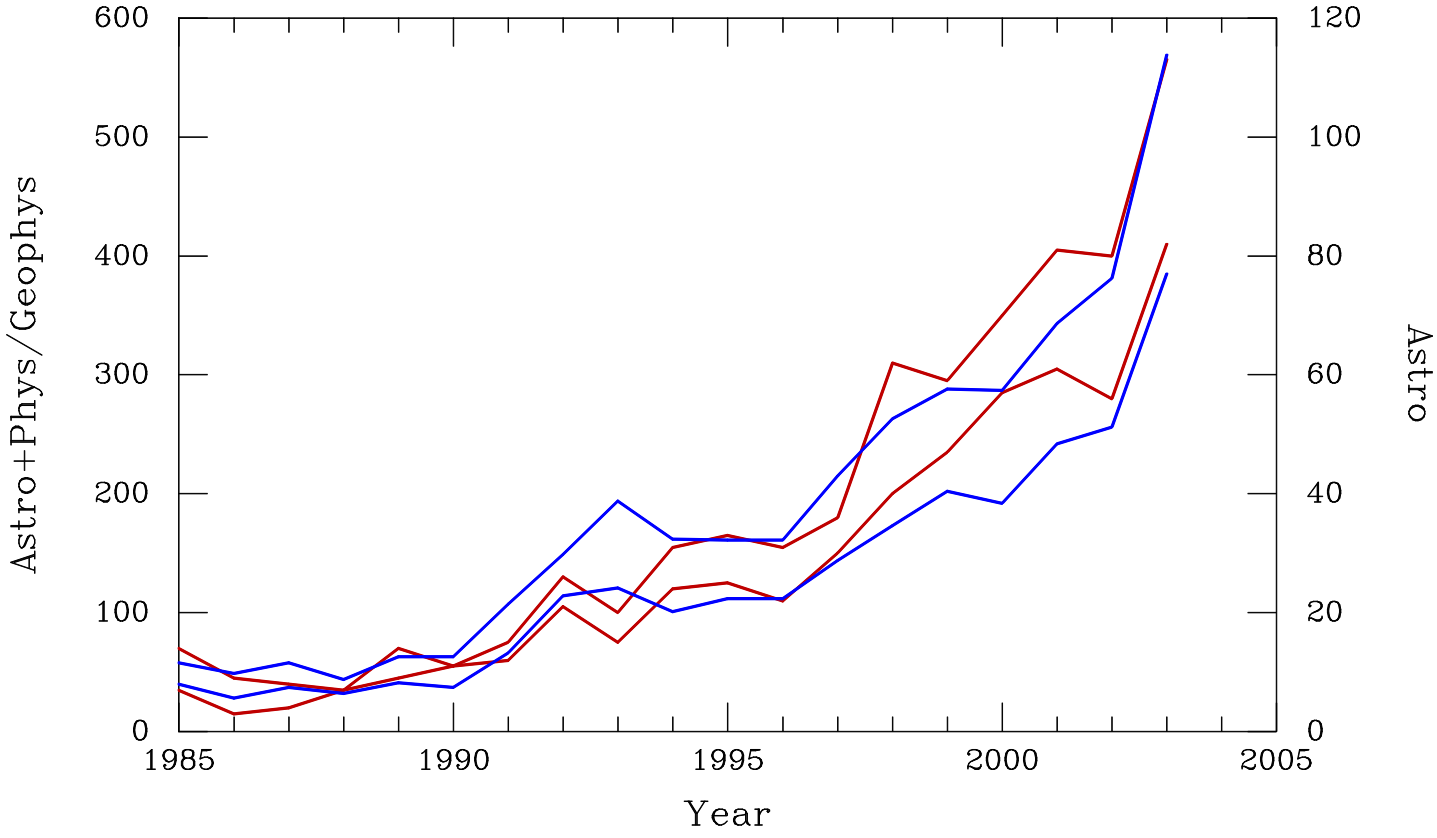
A different approach to *all* statistical inference problems (i.e., not just another method in the list: BLUE, maximum likelihood, χ^2 testing, ANOVA, survival analysis . . .)

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 the *structure of models* rather than *properties of procedures*

An Emerging Subdiscipline

Bayesian papers in Astronomy and
Astronomy, Physics & Geophysics (NASA ADS)



Lower: “Bayes” or “Bayesian” in title or abstract

Upper: Also use of Bayesian terminology in abstract

Note: Several showcase publications aren’t counted!

Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Logic—Some Essentials

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

Build arguments with propositions and logical 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 follows from $\mathcal{P}$

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

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. Statistical methods address validity, but there is no formal approach for addressing factual correctness.

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, *Intro 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
$H|\mathcal{P}$ 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 valid, but may still be rational (likely, plausible, probable)

We seek a quantification of the strength of inductive arguments; we will use the mathematics of deduction to guide us.

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

Integer Representation of Deduction

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

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

Then deduction can be reduced to integer multiplication and addition:

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

Real Number Representation of Induction

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

$$P = 0 \rightarrow \text{Argument is } \textit{invalid}$$

$$= 1 \rightarrow \text{Argument is } \textit{valid}$$

$$\in (0,1) \rightarrow \text{Degree of implication}$$

A mathematical model for induction:

'AND' (product rule) $P(A,B|\mathcal{P}) = P(A|\mathcal{P})\,P(B|A,\mathcal{P})$

$$= P(B|\mathcal{P})\,P(A|B,\mathcal{P})$$

'OR' (sum rule) $P(A \vee B|\mathcal{P}) = P(A|\mathcal{P}) + P(B|\mathcal{P})$

$$-P(A,B|\mathcal{P})$$

We will explore the implications of this model.

The Product Rule

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

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

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. Today many different lines of argument *derive* those rules from various simple and appealing requirements (logical consistency, optimal decisions, conservation of information).

Interpreting Abstract 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 habitiaully 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 $= 273$K Boiling hot $= 373$K
uncertainty	Probability, P	Certainty $= 0$, 1 $p = 1/36$: plausible as "snake's eyes" $p = 1/1024$: plausible as 10 heads

Hypotheses, Data, and Models

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

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

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

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

Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Three Important Theorems

Bayes's Theorem (BT)

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

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

Solve for the *posterior probability*:

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

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

posterior $\propto$ *prior* $\times$ *likelihood*

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

Law of Total Probability (LTP)

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

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

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

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

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

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

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

(*aka* "marginal likelihood")

Normalization

For *exclusive, exhaustive* H_i,

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

Bayesian Inference

Dennis Lindley, "What is a Bayesian?"

> "... a Bayesian is one who holds that the only sensible measure of uncertainty is probability. Or to express the same idea differently and more operationally,
>
> *statements of uncertainty should combine according to the rules of the probability calculus*."

Bayesian inference consists of reporting probabilities for things we are uncertain of.

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

For practical purposes, take the "Three Theorems" to be the "axioms" for the theory.

The Rules in Plain English

With corollaries

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

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

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

Terminology: Likelihood Function

Data enter inferences via $P(D_{\text{obs}}|H_i, I)$.

To assign or calculate this, we'll have to consider what other data, D_k, we might have obtained.

Sampling distribution for data: Dependence on D_k

$$f(k) = P(D_k|H_i, I)$$

Consider this vs. k for *fixed* i. It is normalized over k.

Likelihood for hypothesis: Dependence on H_i

$$\mathcal{L}_i = P(D_{\text{obs}}|H_i, I)$$

Consider this vs. i for D_k *fixed at* D_{obs}. It is *not* a probability for $H_i|\ldots$, and need not be normalized over i.

Continuous Hypothesis Spaces

For hypotheses labeled by a continuous parameter, θ, consider statements about *intervals* of θ:

$$P(\theta \in [\theta_1, \theta_2]|I)$$

Probability density function (pdf):

$$p(\theta) \equiv \lim_{\delta\theta \to 0} \frac{P(\theta \in [\theta, \theta + \delta\theta])}{\delta\theta} \qquad || \; I$$

Statements about intervals $\rightarrow$ inegral of $p(\theta|I)$ over interval.
All the rules hold for pdfs ($\delta\theta$ cancels).

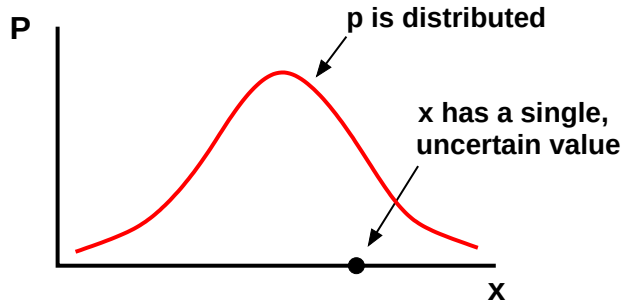
Abuse of notation: We will often conflate $p(\theta|I)$ (a function of the real number θ) with $P(\theta \in [\theta, \theta + \delta\theta]|I)$ (a function of an argument comprised of propositions).

Note the *Skilling conditional:* $|| \; I$

A Bit More On Interpretation

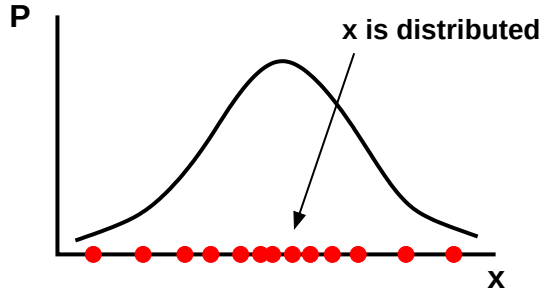
Bayesian

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 an ensemble. $p(x)$ describes variability, how the *values of* x are distributed among the cases in the ensemble:



Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Basic Application: Changing Variables

We know $p(\theta|I)$. What is $p(\phi|I)$ for $\phi = \Phi(\theta)$?

$$\begin{aligned} p(\phi) &= \int d\theta \; p(\phi, \theta) \\ &= \int d\theta \; p(\phi|\theta) \, p(\theta) \qquad\qquad || \; I \end{aligned}$$

First factor is a δ-function:

$$\begin{aligned} p(\phi|\theta, I) &= \delta[\phi - \Phi(\theta)] \\ &= \delta[\theta - \Theta(\phi)] \frac{d\theta}{d\phi} \quad \text{(sum over roots)} \end{aligned}$$

Thus

$$p(\phi|I) = p(\theta|I) \frac{d\theta}{d\phi} \quad \text{(sum over roots)}$$

Basic Application: Propagate Uncertainty

We know $p(\theta, \phi|I)$. What is $p(f|I)$ for $f = F(\theta, \phi)$?

$$\begin{aligned} p(f) &= \int d\theta d\phi \; p(f, \theta, \phi) \\ &= \int d\theta d\phi \; p(f|\theta, \phi) \, p(\theta, \phi) \qquad\qquad || \; I \end{aligned}$$

First factor is a δ-function; change it to one of θ, ϕ:

$$\begin{aligned} p(f|\theta, \phi, I) &= \delta[f - F(\theta, \phi)] \\ &= \delta[\theta - \Theta(f, \phi)] \frac{d\theta}{df} \end{aligned}$$

For example, if $p(\theta, \phi|I) = h(\theta)g(\phi)$,

$$p(f|I) = \int d\phi \; h[\Theta(f, \phi)] \; g(\phi) \; \frac{d\theta}{df}$$

Example: $f = A\theta + B\phi$, h and g are Gaussians at $\hat{\theta}$ and $\hat{\phi}$ with widths σ_θ, σ_ϕ. Let $N(\cdot)$ be standard normal:

$$p(f|I) = \int d\phi \; N\left[\frac{\frac{f-B\phi}{A} - \hat{\theta}}{\sigma_\theta}\right] \times N\left[\frac{\phi - \hat{\phi}}{\sigma_\phi}\right] \times \frac{1}{A}$$

Sum Gaussian exponents, complete square in ϕ, integrate $\rightarrow$ $p(f|I)$ is Gaussian,

$$f = (A\hat{\theta} + B\hat{\phi}) \pm (A^2\sigma_\theta^2 + B^2\sigma_\phi^2)^{1/2}$$

This procedure isn't limited to small uncertainties, or Gaussians; it's completely general.

Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Inference With Parametric Models

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

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

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

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

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

Parameter Estimation

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

Model Uncertainty

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

Hybrid Uncertainty

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

Nuisance Parameters and Marginalization

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

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

What do the data tell us about s?

Marginal posterior distribution

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

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

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

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

Posterior probability for a model:

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

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

Likelihood for model $=$ Average likelihood for its parameters

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

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

Odds and Bayes factors

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

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

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

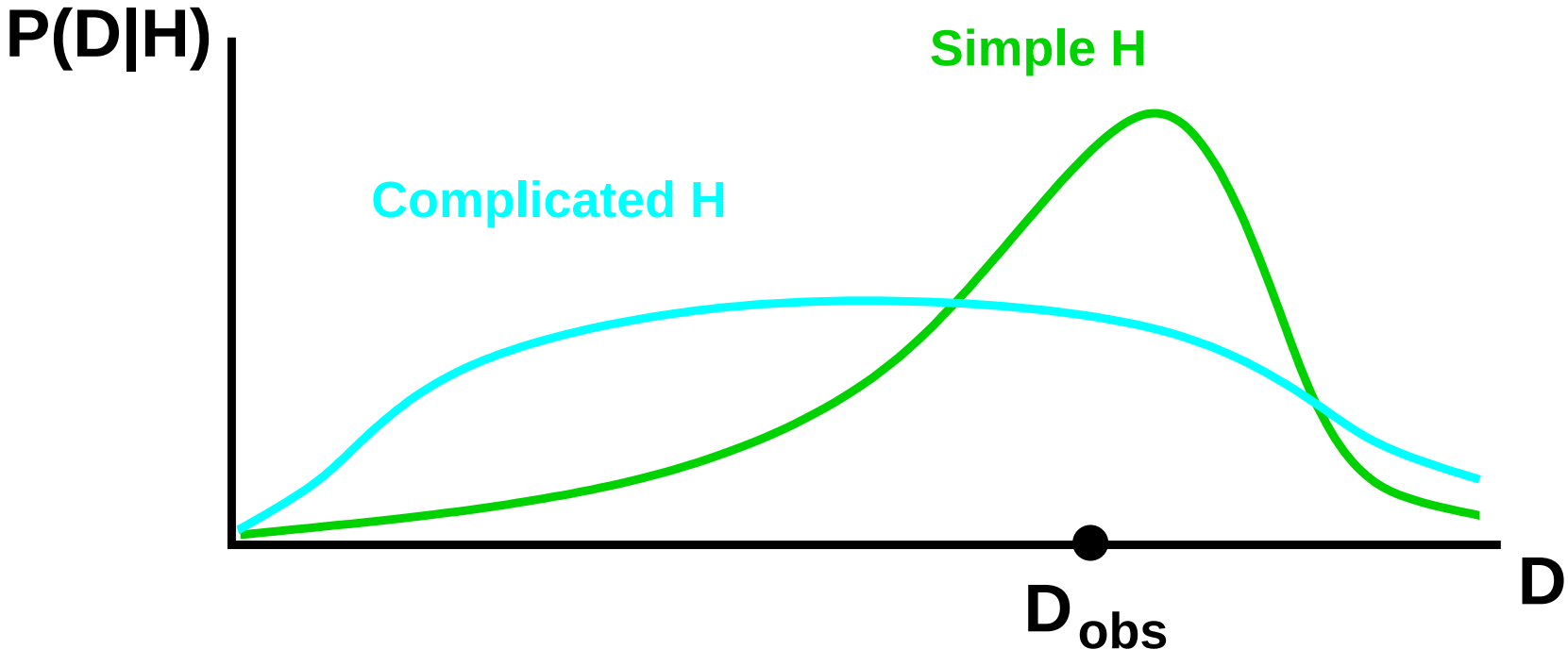
$$B_{ij} \equiv \frac{p(D|M_j,I)}{p(D|M_j,I)}$$

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

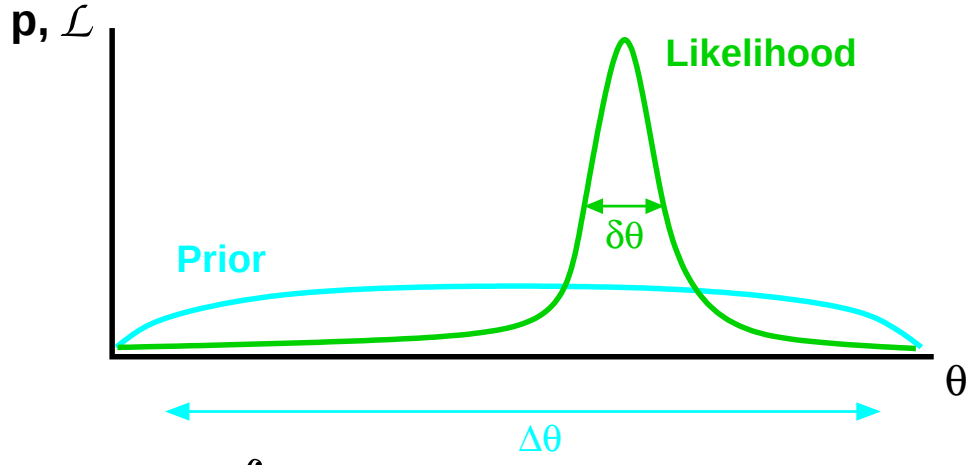
An Automatic Occam's Razor

Predictive probabilities can favor simpler models:

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



The Occam Factor:



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

$$\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i}$$

$$= \text{Maximum Likelihood} \times \text{Occam Factor}$$

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

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

Theme: Parameter Space Volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

- 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 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"

  $\longrightarrow$ Accounts for size of hypothesis space

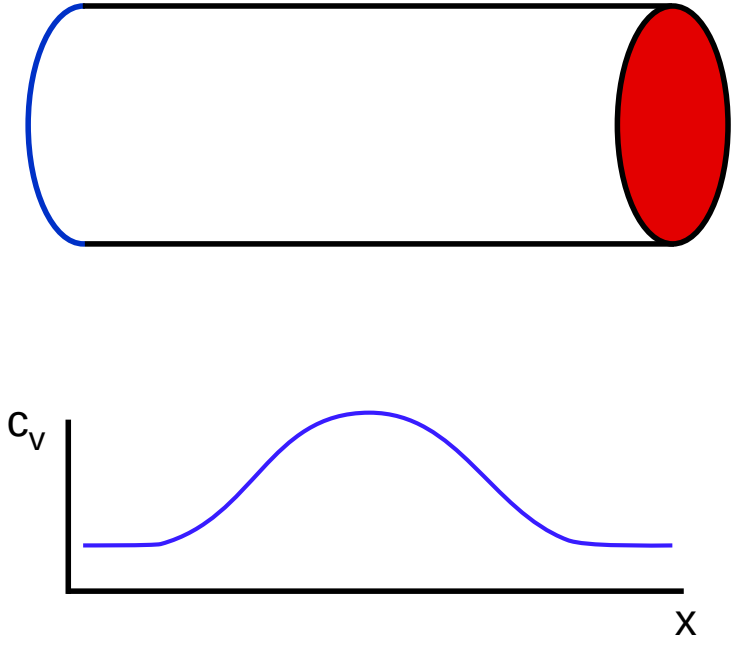
Physical analogy

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

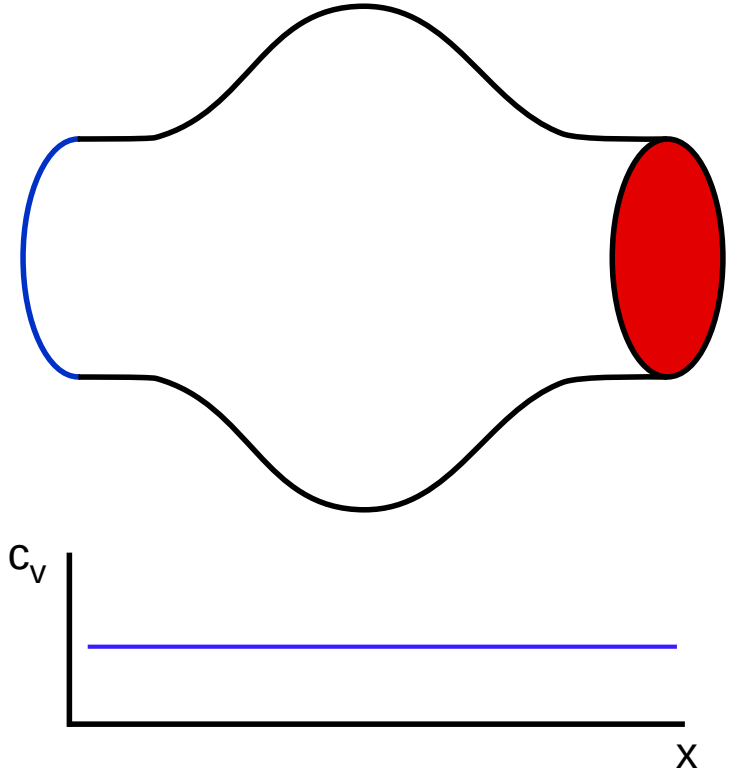
$$\text{Probability:} \quad P \propto \int d\theta \; p(\theta|I) \mathcal{L}(\theta)$$

Max. likelihood focuses on the "hottest" hypotheses.
Bayes focuses on the hypotheses with the most "heat."
Region can have more "heat" due to larger c_v/prior *or* larger volume.

Two bodies for which the most heat is *not*
in the regions of highest temperature



Middle dominates ← nonuniform c_v



Middle dominates ← nonuniform geometry

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 background information. 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").

Lecture 1

- Big picture: The role of statistical inference
- Foundations: Quantifying uncertainty with probability
- Fundamentals: Three important theorems
- Basic applications of probability theory
- Inference with parametric models: Overview
- Inference from binary outcomes

Inference From Binary Outcomes

Parameter Estimation

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

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

$D =$ Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood:

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

Prior:

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

- Intuition: Don't prefer any a 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:

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

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

Prior Predictive:

$$
\begin{aligned}
p(D|M) &= \int da\; a^n (1-a)^{N-n} \\
&= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!}
\end{aligned}
$$

A *Beta integral*, $B(a,b) \equiv \int dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Posterior:

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

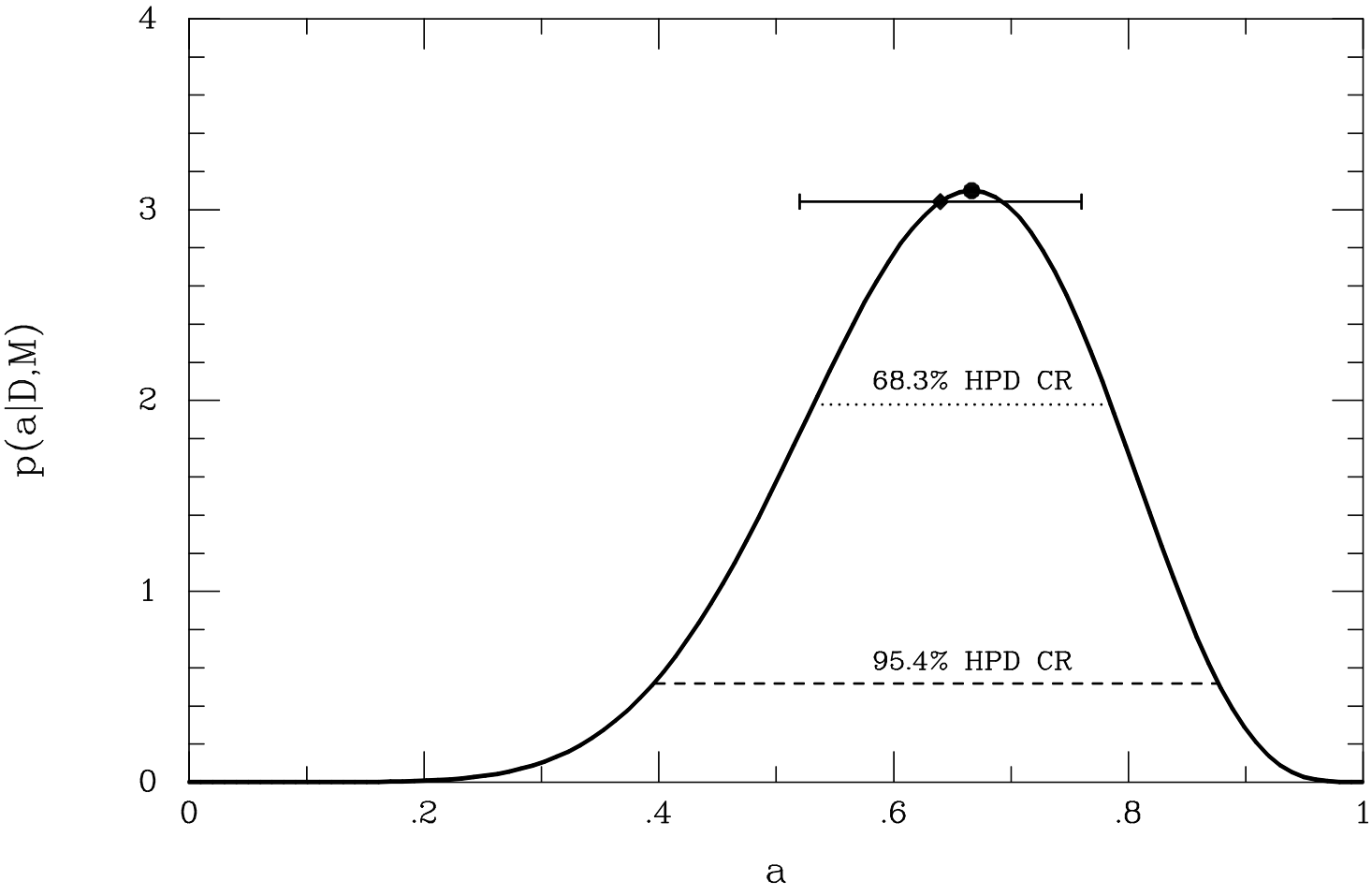
A *Beta distribution*. Summaries:

- Best-fit: $\hat{a} = \frac{n}{N} = 2/3$; $\langle a \rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: $\sigma_a = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$

  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*.

Posterior Distribution



Model Comparison: Equal Probabilities?

M_1: $a = 1/2$

M_2: $a \in [0, 1]$ with flat prior.

Maximum Likelihoods:

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

$$M_2 : \quad \mathcal{L}(2/3) \;=\; \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{p(D|\hat{a}, M_2)} = 0.51$$

Maximum likelihoods favor M_2 (failures more probable).

Bayes Factor (ratio of model likelihoods):

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

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

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.

Binary Outcomes: Binomial Distribution

Suppose $D = n$ (number of heads in N trials), rather than the actual sequence. What is $p(a|n, M)$?

Likelihood:

Let $S =$ a sequence of flips with n heads.

$$\begin{aligned} p(n|a, M) &= \sum_S p(S|a, M)\, p(n|S, a, M) \\ &= a^n (a - a)^{N-n} C_{n,N} \end{aligned}$$

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

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

The *binomial distribution* for n given a, N.

Posterior:

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

$$\begin{aligned} p(n|M) &= \frac{N!}{n!(N-n)!}\int da\; a^n(1-a)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

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

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 predifined number of successes, $n = 8$. What is $p(a|N, M)$?

Likelihood:

$p(N|a, M)$ is probability for $n - 1$ successes in $N - 1$ trials, times probability that the final trial is a success: Let $S =$ a sequence of flips with n heads.

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

The *negative binomial distribution* for N given a, n.

Posterior:

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

$$p(D|M) \;\; = \;\; C'_{n,N}\int da\; a^n(1-a)^{N-n}$$

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

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 $p(a|D, M)$?

Likelihood:

$p(D|a, M)$ is the binomial distribution times the probability that the weather allowed N samples, $W(N)$:

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

Let $C_{n,N} = W(N)\binom{N}{n}$. We get the same result as before!

Likelihood Principle

To assign $\mathcal{L}(H_i) = p(D_{\text{obs}}|H_i, I)$, 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 thus depend on details of the sample space that may be irrelevant in a Bayesian calculation.

Key Ideas

- Statistics as hypothesis appraisal—one part of data analysis
- Probability as quantifying strength of arguments
- Bayesian inference: *All* of probability theory
  - Bayes's theorem
  - Law of total probability
- Changing variables; propagating uncertainty
- Parametric models: Role of parameter space volumes
- Binary outcomes: Sufficiency, likelihood principle