Introduction to Bayesian inference: Fundamentals

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

IAC Winter School, 3–4 Nov 2014

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!

> *"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

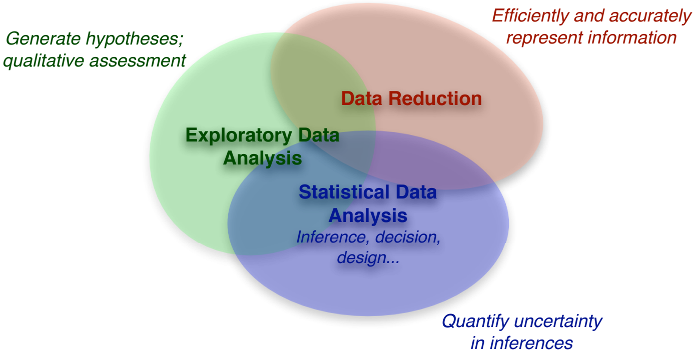
We gather data so they may speak for or against existing hypotheses, and guide the formation of new hypotheses

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

Data analysis

Building & Appraising Arguments Using Data

Modes of Data Analysis


Generate hypotheses; qualitative assessment
Efficiently and accurately represent information
Data Reduction
Exploratory Data Analysis
Statistical Data Analysis
Inference, decision, design...
Quantify uncertainty in inferences

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/χ^2 minimization, maximum likelihood, ANOVA, survival analysis . . .)

- Focuses on *deriving consequences of modeling assumptions* rather than *devising and calibrating procedures*

Frequentist vs. Bayesian statements

"The data D_{obs} support conclusion C . . . "

Frequentist assessment

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

Probabilities are properties of *procedures*, not of particular results

Bayesian assessment

"The strength of the chain of reasoning from the model and D_{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)

Fundamentals

❶ **Confidence intervals vs. credible intervals**

❷ **Foundations: Logic & probability theory**

❸ **Probability theory for data analysis: Three theorems**

❹ **Inference with parametric models**

Parameter Estimation

Model Uncertainty

Fundamentals

❶ **Confidence intervals vs. credible intervals**

❷ **Foundations: Logic & probability theory**

❸ **Probability theory for data analysis: Three theorems**

❹ **Inference with parametric models**
- Parameter Estimation
- Model Uncertainty

A Simple (?) confidence region

Problem

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

Report a *point estimate*, and a *region* summarizing the uncertainty

Model

$$\begin{aligned} p(x_i|\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ \text{Equivalently,}\quad x_i &\sim \mathcal{N}(\mu,\sigma^2) \end{aligned}$$

Here assume σ is *known*; we are uncertain about μ

Classes of variables

- μ is the unknown we seek to estimate—the *parameter*. The *parameter space* is the space of possible values of μ—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

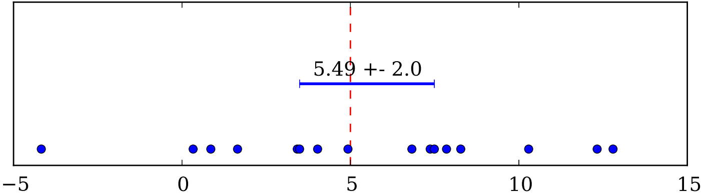
Let $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

- "Standard error" (rms error) is $\sigma/\sqrt{N}$
- "1σ" interval: $\bar{x} \pm \sigma/\sqrt{N}$ with conf. level CL $= 68.3\%$
- "2σ" 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 $\sigma/\sqrt{N} = 1$

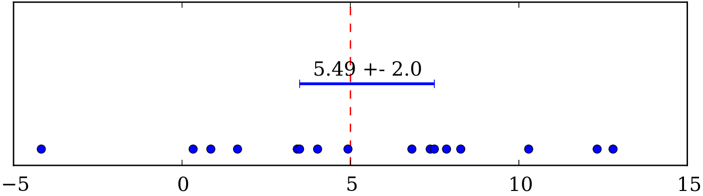
What is the CL associated with this interval?



Some simulated data

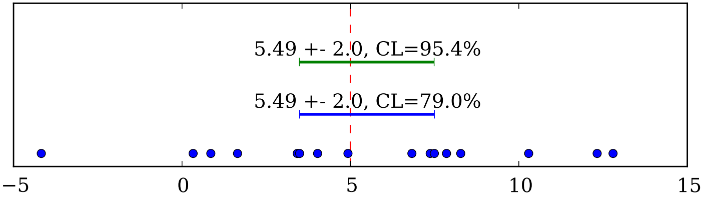
Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$

What is the CL associated with this interval?



The (frequentist) confidence level for this interval is 79.0%

Two intervals



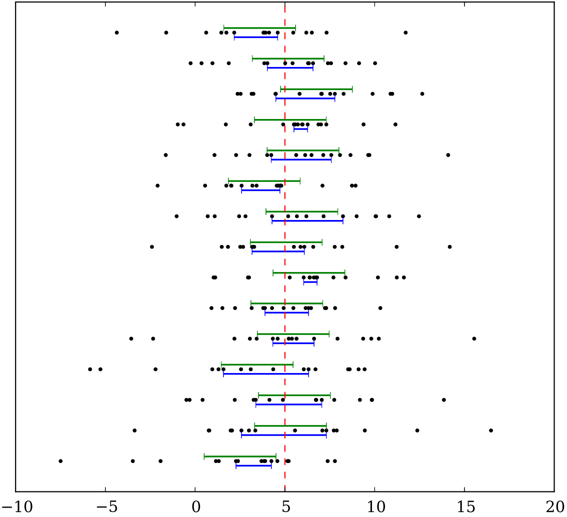
- Green interval: $\bar{x} \pm 2\sigma/\sqrt{N}$
- Blue interval: Let $x_{(k)} \equiv k$'th order statistic
  Report $[x_{(6)}, x_{(11)}]$ (i.e., leave out 5 outermost each side)

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, with

$$x_i \sim N(\mu, 1)$$

We want to estimate μ, including some *quantification of uncertainty* in the estimate: an interval *with a probability attached*

Frequentist approaches: method of moments, BLUE, least-squares/χ^2, maximum likelihood

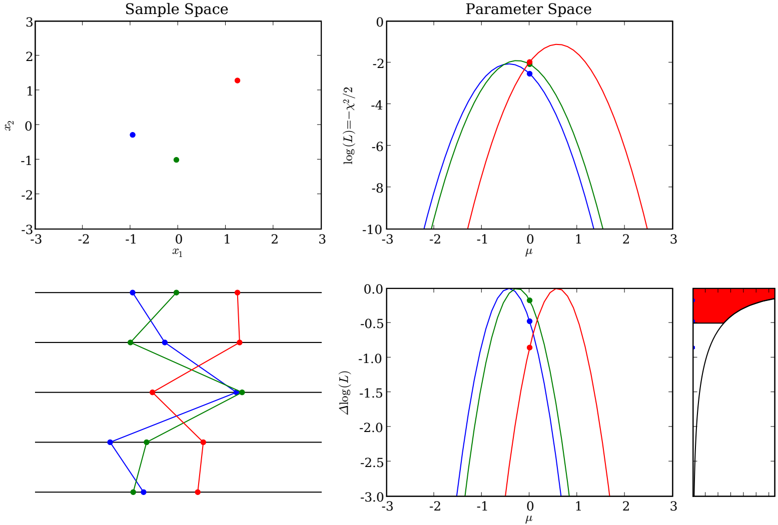
Focus on likelihood (equivalent to χ^2 here); this is closest to Bayes:

$$\begin{aligned} \mathcal{L}(\mu) &= p(\{x_i\}|\mu) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

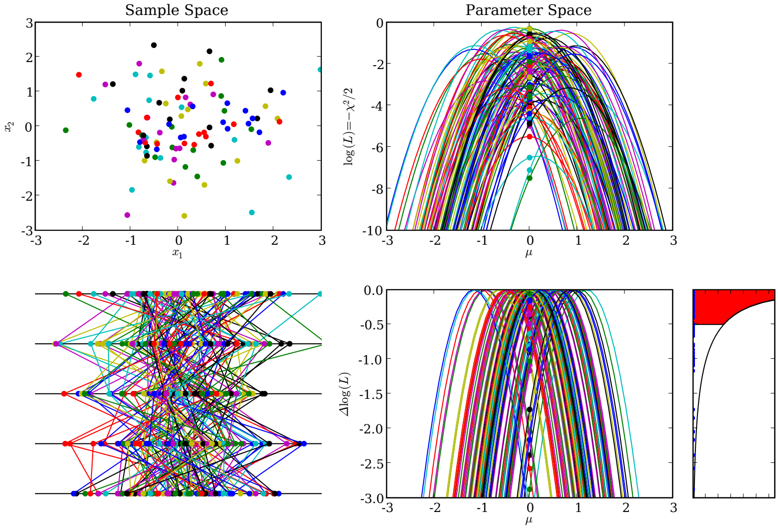
Estimate μ from maximum likelihood (minimum χ^2)
Define an interval and its coverage frequency from the $\mathcal{L}(\mu)$ curve

Construct an interval procedure for known μ

Likelihoods for 3 simulated data sets, $\mu = 0$

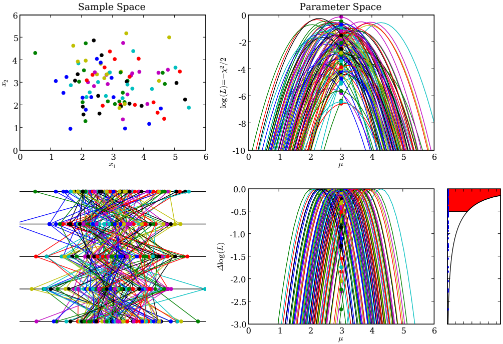


Likelihoods for 100 simulated data sets, $\mu = 0$



Explore dependence on μ

Likelihoods for 100 simulated data sets, $\mu = 3$

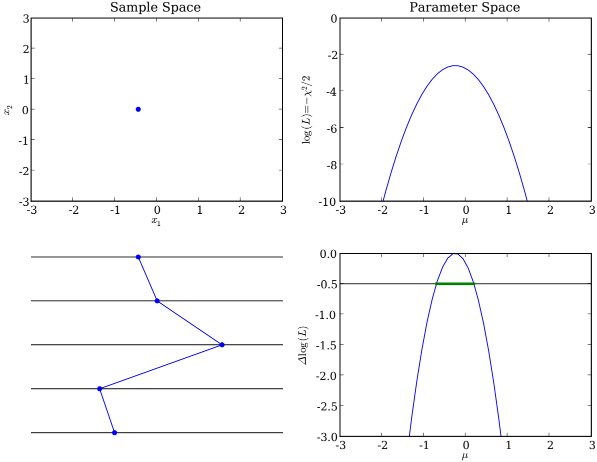


Luckily the $\Delta \log \mathcal{L}$ distribution is the same!
($\Delta \log \mathcal{L}$ is a *pivotal quantity*)

If it weren't, define *confidence level* $=$ maximum coverage over all μ (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 (red 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 μ

Convert likelihood to a probability distribution over μ via *Bayes's theorem*:

$$
\begin{aligned}
p(A, B) &= p(A)p(B|A) \\
&= p(B)p(A|B) \\
\to p(A|B) &= p(A)\frac{p(B|A)}{p(B)}, \quad \text{Bayes's th.}
\end{aligned}
$$

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

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

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 wrt some reparameterization motivated by a symmetry in the problem)

Gaussian problem posterior distribution

For the Gaussian example, a bit of algebra ("complete the square") gives:

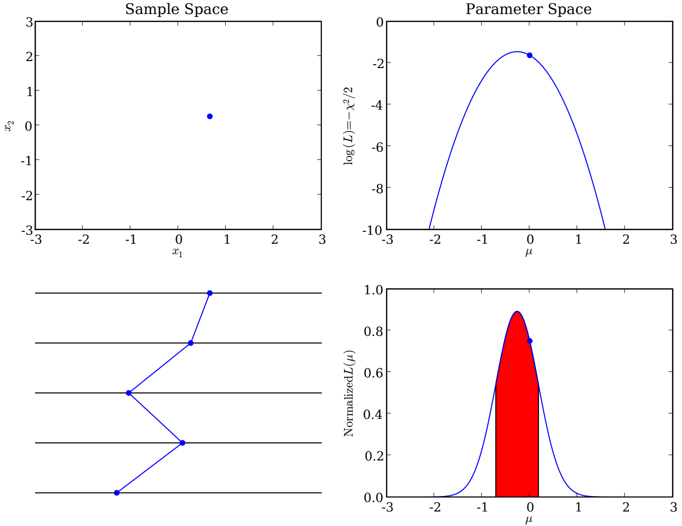
$$\begin{aligned}\mathcal{L}(\mu) &\propto \prod_i \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &\propto \exp\left[-\frac{(\mu-\bar{x})^2}{2(\sigma/\sqrt{N})^2}\right]\end{aligned}$$

The likelihood is Gaussian in μ

Flat prior → posterior density for μ 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



Posterior summaries

- Posterior mean is $\langle\mu\rangle \equiv \int d\mu\, \mu\, p(\mu|D_{\rm obs}) = \bar{x}$
- 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*:

$$\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} d\mu\; p(\mu|D_{\rm obs}) \approx 0.683$$

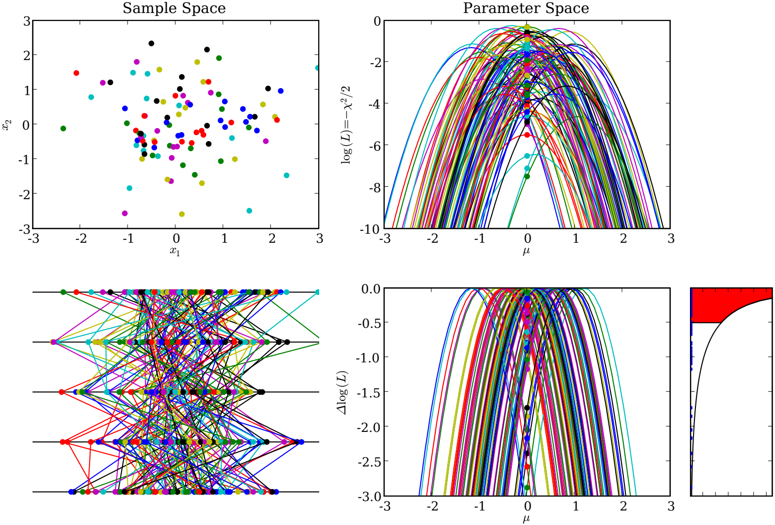
- $\bar{x} \pm 2\sigma/\sqrt{N}$ is a 95.4% credible region

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 *coincidence*!

When might results differ? ($\mathcal{F}$ = frequentist, $\mathcal{B}$ = Bayes)

- If $\mathcal{F}$ procedure doesn't use likelihood directly
- If $\mathcal{F}$ 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, corrected profile likelihood, conditional inference)
- If $\mathcal{B}$ uses important prior information

Also, for a different task—comparison of parametric models—the approaches are qualitatively different (significance tests & info criteria vs. Bayes factors)

Supplement — Multivariate confidence and credible regions: *parametric bootstrapping* vs. *posterior sampling*

Bayesian and Frequentist inference

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. . . .

Roderick Little, ASA President's Address (2005)

> 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 arXiv:1208.3035 [by TL] for discussion/references)

Fundamentals

1. **Confidence intervals vs. credible intervals**

2. **Foundations: Logic & probability theory**

3. **Probability theory for data analysis: Three theorems**

4. **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:

$$
\begin{aligned}
H|\mathcal{P}: \quad & \text{Premises } \mathcal{P} \text{ imply } H \\
& H \text{ may be deduced from } \mathcal{P} \\
& H \text{ follows from } \mathcal{P} \\
& H \text{ is true given that } \mathcal{P} \text{ is true}
\end{aligned}
$$

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) 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.

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 premises 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 premises 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* → chains of discovery

Every argument has a set of premises defining a fixed *context* in which the argument is 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

Representing induction with $[0,1]$ calculus

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

$$\begin{aligned} P &= 1 \rightarrow \text{Argument is \textit{deductively valid}} \\ &= 0 \rightarrow \text{Premises imply } \overline{H} \\ &\in (0,1) \rightarrow \text{Degree of deducibility} \end{aligned}$$

Mathematical model for induction

$$\text{`AND' (product rule):} \quad P(A \wedge B|\mathcal{P}) = P(A|\mathcal{P})\, P(B|A \wedge \mathcal{P}) = P(B|\mathcal{P})\, P(A|B \wedge \mathcal{P})$$

$$\text{`OR' (sum rule):} \quad P(A \vee B|\mathcal{P}) = P(A|\mathcal{P}) + P(B|\mathcal{P}) - P(A \wedge B|\mathcal{P})$$

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

Firm foundations

Many different formal 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)

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."

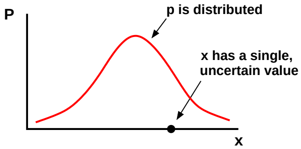
A Thermal Analogy

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

Interpreting PDFs

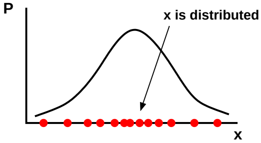
Bayesian

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:



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:



Fundamentals

❶ **Confidence intervals vs. credible intervals**

❷ **Foundations: Logic & probability theory**

❸ **Probability theory for data analysis: Three theorems**

❹ **Inference with parametric models**
- Parameter Estimation
- Model Uncertainty

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):
$$\begin{aligned} P(H_1 \vee H_2|I) &= P(H_1|I) + P(H_2|I) \\ &\quad - P(H_1, H_2|I) \end{aligned}$$

'AND' (product rule):
$$\begin{aligned} P(H_1, D_{\rm obs}|I) &= P(H_1|I)\, P(D_{\rm obs}|H_1, I) \\ &= P(D_{\rm obs}|I)\, P(H_1|D_{\rm obs}, I) \end{aligned}$$

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

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 $\propto$ *prior* $\times$ *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 $A = D_{\rm obs}$, $B_i = H_i$; then

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

prior predictive for $D_{\rm obs}$ = 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 context. 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")

Fundamentals

1. **Confidence intervals vs. credible intervals**

2. **Foundations: Logic & probability theory**

3. **Probability theory for data analysis: Three theorems**

4. **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*:

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

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: $D = D_{obs}$

Classes of Problems

Single-model inference

Premise = choice of single model (specific i)

Parameter estimation: What can we say about θ_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*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
  What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D', accounting for model uncertainty?

Model checking

Premise = $M_1 \vee$ "all" alternatives

Is M_1 adequate? (predictive tests, calibration, robustness)

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

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

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

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 δb_s

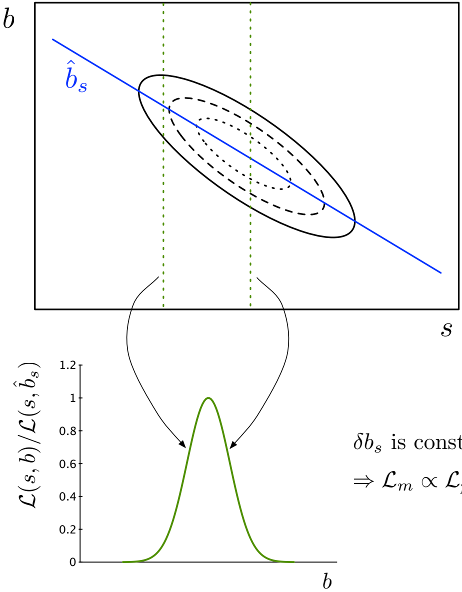
$$\begin{aligned}
\mathcal{L}_m(s) &\equiv \int db\, p(b|s)\, \mathcal{L}(s,b) \\
&\approx p(\hat{b}_s|s)\, \mathcal{L}(s, \overset{\text{best } b \text{ given } s}{\hat{b}_s})\, \underset{b \text{ uncertainty given } s}{\delta b_s}
\end{aligned}$$

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

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*

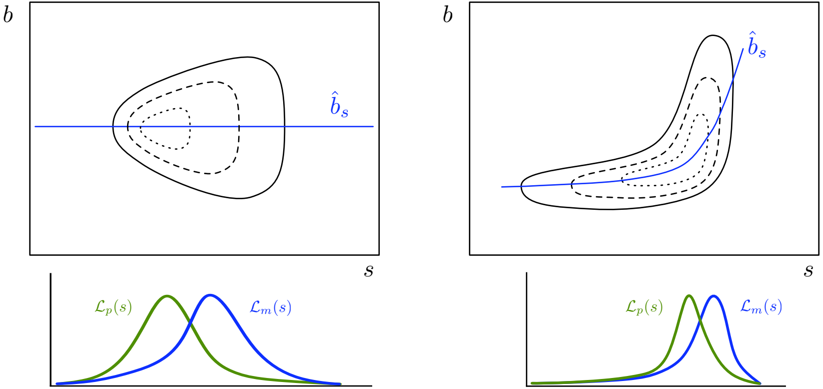
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



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 $F = f(\theta)$?

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

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

$I = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models

$H_i = M_i$ — Hypothesis chooses a model

Posterior probability for a model

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

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

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*:

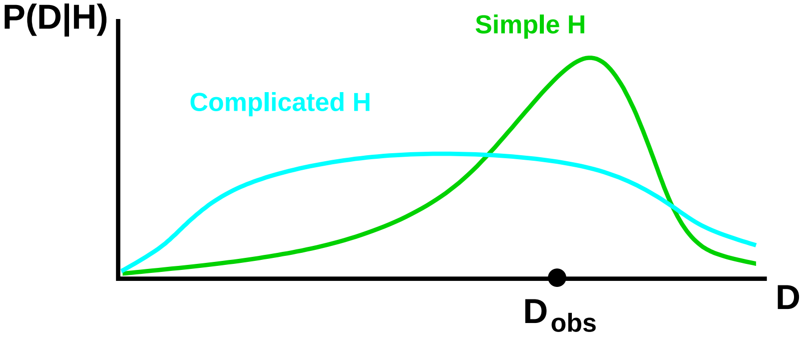
$$B_{ij} \equiv \frac{p(D|M_i,I)}{p(D|M_j,I)}$$

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

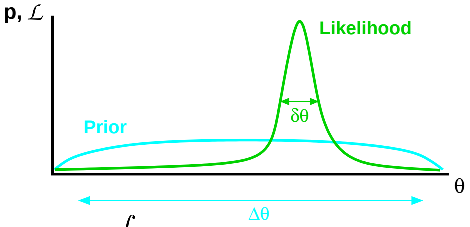
An Automatic Ockham's Razor

Predictive probabilities can favor simpler models

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



The Ockham Factor



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

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

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)

- 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

$$\begin{aligned} \text{Heat} \quad Q &= \int d\mathbf{r}\, [\rho(\mathbf{r}) c_v(\mathbf{r})]\, T(\mathbf{r}) \\ \text{Probability} \quad P &\propto \int d\theta\; p(\theta)\mathcal{L}(\theta) \end{aligned}$$

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

Recap of Key Ideas

Probability as generalized logic

Probability quantifies the *relative 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

$$p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) = \sum p(\text{Hypothes}\underline{\textbf{is}} \mid \text{Data})$$

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