From least squares to multilevel modeling:
A graphical introduction to Bayesian inference

Tom Loredo
Cornell Center for Astrophysics and Planetary Science

Session site:

AAS 227 — 6 Jan 2016
A Simple (?) confidence region

**Problem**

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

**Model**

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

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

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

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

Standard inferences

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

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

Consider a case with $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$

Simulate data with true $\mu = 5$

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

Consider a case with $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$

Simulate data with true $\mu = 5$

What is the CL associated with this interval?

The confidence level for this interval is $79.0\%$. 

![Graph showing simulated data and confidence interval]
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
Probabilities for procedures vs. arguments

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

Frequentist assessment

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

Probability is a property of a procedure, not of a particular result

Procedure specification relies on the ingenuity/experience of the analyst
“The data $D_{\text{obs}}$ support conclusion $C$ . . . ”

**Bayesian assessment**

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

Probability is a property of an *argument*: a statement that a hypothesis is supported by *specific, observed data*

The function of the data to be used is uniquely specified by the model

Long-run performance must be separately evaluated (and is typically good by frequentist criteria)
Bayesian statistical inference

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

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

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

1 Probability: variability vs. argument strength

2 Computation: mock data vs. mock hypotheses
   - Confidence vs. credible regions
   - Posterior sampling
   - Nuisance parameters & marginalization

3 Graphical models: mock data and mock hypotheses
Agenda

1. **Probability: variability vs. argument strength**

2. **Computation: mock data vs. mock hypotheses**
   - Confidence vs. credible regions
   - Posterior sampling
   - Nuisance parameters & marginalization

3. **Graphical models: mock data and mock hypotheses**
Understanding probability

“$X$ is random . . . ”

Frequentist understanding

“The value of $X$ varies across repeated observation or sampling.”

Probability quantifies variability

Bayesian understanding

“The value of $X$ in the case at hand is uncertain.”

Probability measures the strength with which the available information supports possible values for $X$ (before and/or after measurement or observation)
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:
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:
Twiddle notation for the normal distribution

\[ \text{Norm}(x, \mu, \sigma) \equiv \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{\sigma^2} \right] \]

**Frequentist**

random \hspace{1cm} \text{fixed but unknown}

\[ p(x \mid \mu, \sigma) = \text{Norm}(x, \mu, \sigma) \]

\[ x \sim N(\mu, \sigma^2) \]

“\(x\) is distributed as normal with mean…”

**Bayesian**

random \hspace{1cm} \text{random or known}

\[ p(x \mid \mu, \sigma) = \text{Norm}(x, \mu, \sigma) \]

\[ x \sim N(\mu, \sigma^2) \]

“The probability for \(x\) is distributed as normal with mean…”
Agenda

1 Probability: variability vs. argument strength

2 Computation: mock data vs. mock hypotheses
   - Confidence vs. credible regions
   - Posterior sampling
   - Nuisance parameters & marginalization

3 Graphical models: mock data and mock hypotheses
Confidence interval for a normal mean

Suppose we have a sample of $N = 5$ values $x_i$,

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

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

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

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

$$\mathcal{L}(\mu) = p(x_i | \mu)$$

$$= \prod_i \frac{1}{\sigma \sqrt{2\pi}} e^{-(x_i - \mu)^2 / 2\sigma^2}; \quad \sigma = 1$$

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

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

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

[Skip some crucial steps here: CL vs. coverage, pivotal quantities...]
Report the green region, with coverage as calculated for ensemble of hypothetical data (green region, previous slide).
Likelihood to probability via Bayes’s theorem

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

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

$$p(A, B) = p(A)p(B|A)$$
$$= p(B)p(A|B)$$
$$\Rightarrow p(A|B) = p(A) \frac{p(B|A)}{p(B)}$$, Bayes’s th.

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

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

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

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

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

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

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

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

Sample Space

Parameter Space

Joint Space

Normalized $L(\mu)$

200 post. samples
Inference as manipulation of the joint distribution

Bayes’s theorem in terms of the joint distribution:

\[ p(\mu) \times p(\vec{x}|\mu) = p(\mu, \vec{x}) = p(\vec{x}) \times p(\mu|\vec{x}) \]

Components of Bayes’s theorem for a problem with a 1-D parameter space (\(\theta\)) and a 2-D sample space (\(y\)), with observed data \(y_d\), and modeling assumptions \(A\)
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, the law of total probability $\rightarrow$ marginalize:

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

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

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

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

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

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

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

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

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

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

Marginalization offers a generalized form of error propagation, without approximation
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*

\[
Q = \int d\vec{r} \left[ \rho(\vec{r}) c_v(\vec{r}) \right] T(\vec{r})
\]

\[
P \propto \int d\theta \ p(\theta) L(\theta)
\]

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

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

A high-\(L\) region may contain little probability if its prior is low or if its volume is small
Agenda

1. **Probability:** variability vs. argument strength

2. **Computation:** mock data vs. mock hypotheses
   - Confidence vs. credible regions
   - Posterior sampling
   - Nuisance parameters & marginalization

3. **Graphical models:** mock data and mock hypotheses
Density estimation with measurement error

*Introduce latent/hidden/incidental parameters*

Suppose \( f(x|\theta) \) is a distribution for an observable, \( x \).

From \( N \) precisely measured samples, \( \{x_i\} \), we can infer \( \theta \) from

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

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

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

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

Graph specifies the form of the *joint distribution*:

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

Posterior from BT: $p(\theta | \{x_i\}) = p(\theta, \{x_i\}) / p(\{x_i\})$
But what if the $x$ data are noisy, $D_i = \{x_i + \epsilon_i\}$?

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

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

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

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

Key point: **Maximizing over $x_i$ and integrating over $x_i$ can give very different results!**
Graphical representation

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

A two-level multi-level model (MLM)
Recap of Key Ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

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

Use *all* of probability theory for this

**Bayes’s theorem**

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

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

**Law of total probability**

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

The support for a *compound/composite* hypothesis must account for all the ways it could be true
Bayesian tutorials (basics & MLMs):
CASt 2015 Summer School
2014 Canary Islands Winter School

Tutorials on Bayesian computation:
SCMA 5 Bayesian Computation tutorial notes
CASt 2014 Supplement Sessions

Literature entry points:
Overview of MLMs in astronomy: arXiv:1208.3036
Discussion of recent B vs. F work: arXiv:1208.3035

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