An Introduction to
Bayesian Data Analysis

Lecture 1

D. S. Sivia
St. John's College
Oxford, England

April 29, 2012

Some recommended books

Data analysis: a Bayesian tutorial
D. S. Sivia (1996), Oxford University Press; with *J. Skilling* (2006)

Bayesian logical data analysis for the physical sciences
P. Gregory (2005), Cambridge University Press

Information theory, inference and learning algorithms
D.J.C. MacKay (2004), Cambridge University Press

Probability theory: the logics of science
E.T. Jaynes (2003), Cambridge University Press

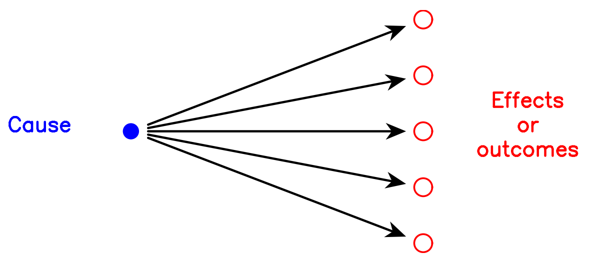
Bayesian reasoning in data analysis
G. D'Agostini (2003), World Scientific Publishing

Outline

- The basics
- Parameter estimation I
- Parameter estimation II
- Model selection
- Assigning probabilities
- Non-parametric estimation
- Experimental design
- Least-squares extensions*
- Nested sampling
- Quantification

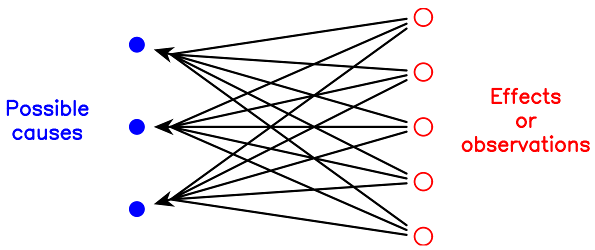
Introduction (1)

(a)



Deductive logic
(pure maths)

(b)



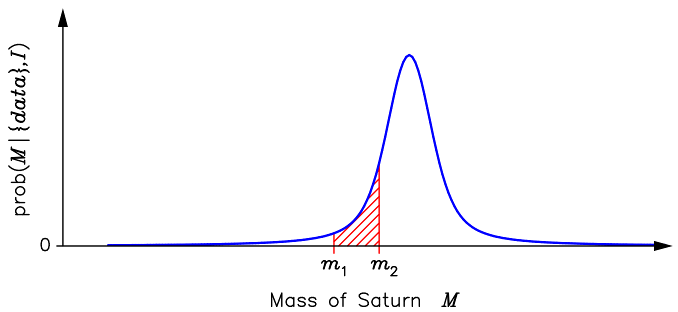
Inductive logic
(science)

or plausible reasoning

Introduction (2)

- Bernoulli (1713), Bayes (1763) and Laplace (1812)

  — developed probability theory to reason in situations where we cannot argue with certainty.



"... bet of 11,000 against 1 that the error of this result is not $\frac{1}{100}$ of its value."

[Laplace]

Introduction (3)

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

[Laplace]

- Cox (1946)

  — showed that any method of plausible reasoning that satisfies simple rules of *logical consistency* must be equivalent to the use of ordinary probability theory.

- Bayesian interpretation of probability
  - A probability encodes a state of knowledge.
  - All probabilities are conditional.

Basic rules and corollaries

- Range : $0 \leqslant \text{prob}(X|I) \leqslant 1$
- Sum rule : $\text{prob}(X|I) + \text{prob}\big(\overline{X}|I\big) = 1$
- Product rule : $\text{prob}(X,Y|I) = \text{prob}(X|Y,I) \times \text{prob}(Y|I)$
- Bayes' theorem : $\text{prob}(X|Y,I) = \dfrac{\text{prob}(Y|X,I) \times \text{prob}(X|I)}{\text{prob}(Y|I)}$
- Marginalisation : $\text{prob}(X|I) = \text{prob}(X,Y|I) + \text{prob}\big(X,\overline{Y}|I\big)$

Basic rules and corollaries

- Range : $0 \leqslant \text{prob}(X|I) \leqslant 1$
- Sum rule : $\text{prob}(X|I) + \text{prob}\big(\overline{X}|I\big) = 1$
- Product rule : $\text{prob}(X,Y|I) = \text{prob}(X|Y,I) \times \text{prob}(Y|I)$

- Bayes' theorem : $\text{prob}(X|Y,I) = \dfrac{\text{prob}(Y|X,I) \times \text{prob}(X|I)}{\text{prob}(Y|I)}$
- Marginalisation : $\text{prob}(X|I) = \sum_{j=1}^{M} \text{prob}(X,Y_j|I)$

Basic rules and corollaries

- Range : $0 \leqslant \text{prob}(X|I) \leqslant 1$
- Sum rule : $\text{prob}(X|I) + \text{prob}\big(\overline{X}|I\big) = 1$
- Product rule : $\text{prob}(X,Y|I) = \text{prob}(X|Y,I) \times \text{prob}(Y|I)$

- Bayes' theorem : $\text{prob}(X|Y,I) = \dfrac{\text{prob}(Y|X,I) \times \text{prob}(X|I)}{\text{prob}(Y|I)}$
- Marginalisation : $\text{prob}(X|I) = \int_{-\infty}^{+\infty} \text{prob}(X,Y|I)\, \mathrm{d}Y$

1-Parameter estimation

Data $\mathbf{D} = r$ heads in n flips. Is this a fair coin?

If H is the probability of getting a head, what is the value of H?

$$\underbrace{\text{prob}(H|\mathbf{D},I)}_{\textit{Posterior}} \propto \underbrace{\text{prob}(\mathbf{D}|H,I)}_{\textit{Likelihood}} \times \underbrace{\text{prob}(H|I)}_{\textit{Prior}} \qquad \text{(Bayes')}$$

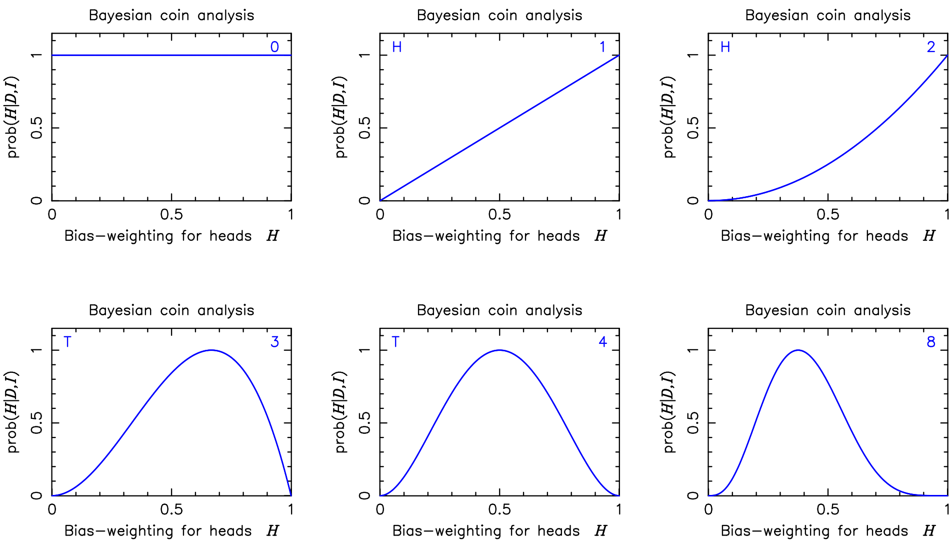
- Binomial likelihood:

$$\text{prob}(\mathbf{D}|H,I) = \frac{n!}{r!\,(n-r)!}\, H^r\,(1-H)^{n-r}$$

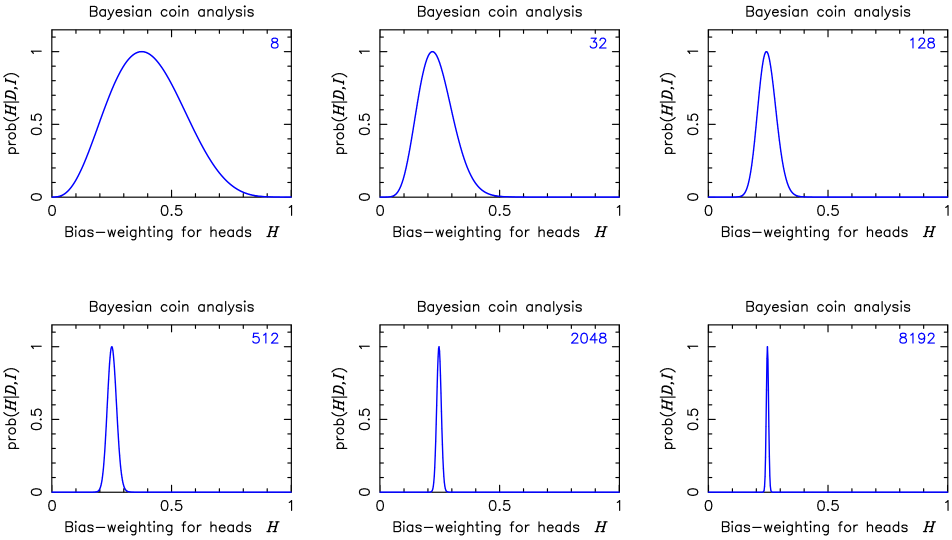
- Ignorant prior:

$$\text{prob}(H|I) = \begin{cases} 1 & \text{for } 0 \leqslant H \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

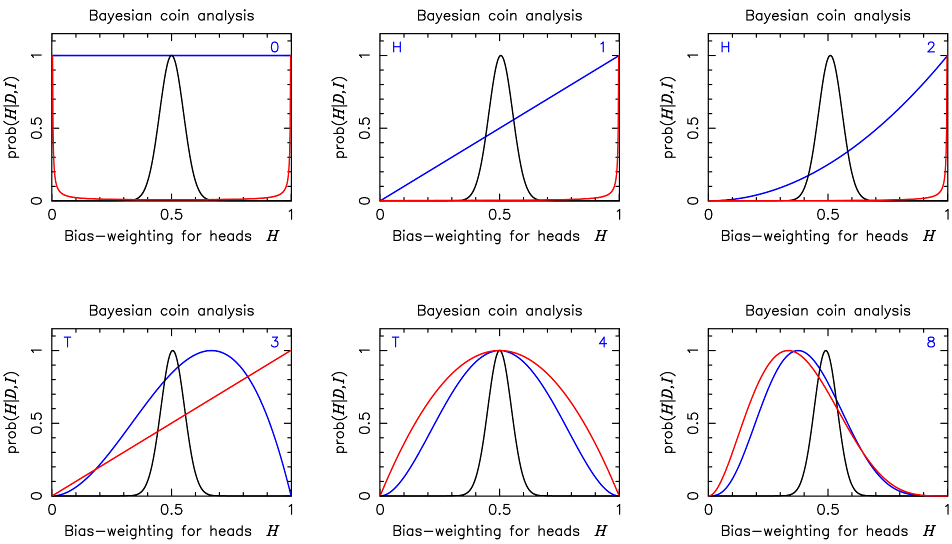
Coin example: uniform prior (1)



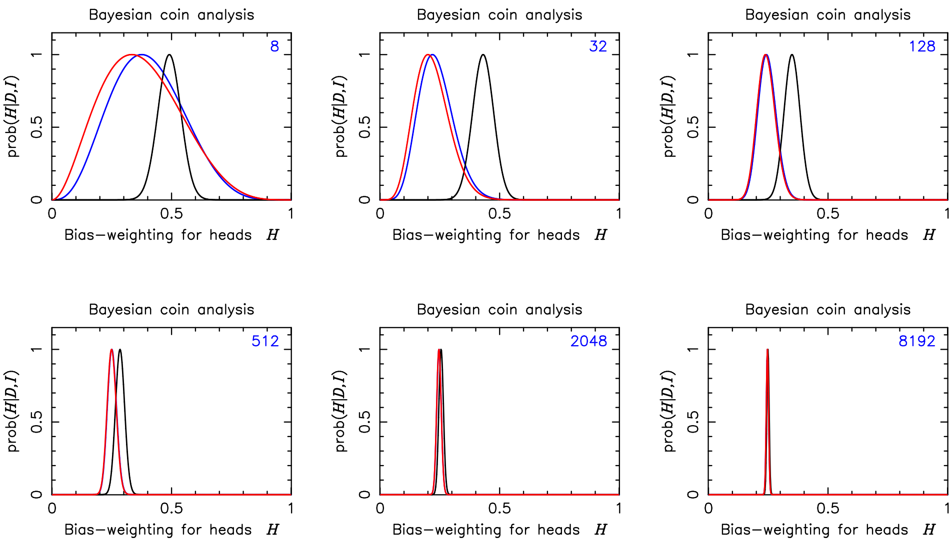
Coin example: uniform prior (2)



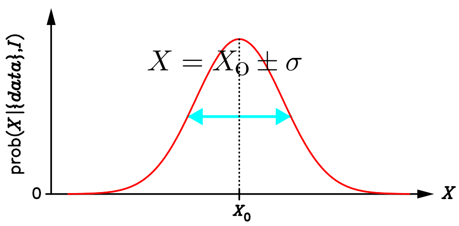
Coin example: alternative priors (1)



Coin example: alternative priors (2)



Summarising the inference

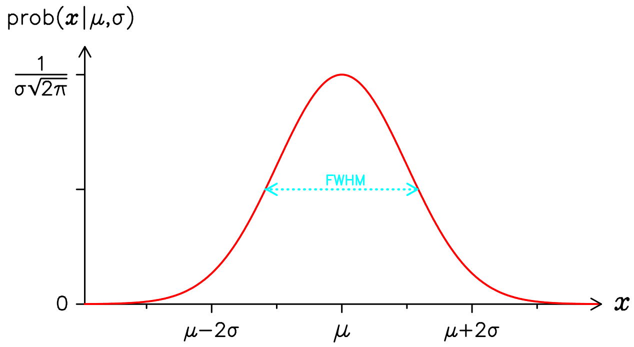


Let $L = \log_e\left[\,\mathrm{prob}(X\,|\,\{data\},I\,)\,\right]$

Taylor: $L(X) = L(X_0) + \frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}X^2}\right|_{X_0}(X-X_0)^2 + \cdots$, where $\left.\frac{\mathrm{d}L}{\mathrm{d}X}\right|_{X_0} = 0$

$$\Rightarrow\ \mathrm{prob}(X\,|\,\{data\},I\,) \approx \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(X-X_0)^2}{2\,\sigma^2}\right],\quad \sigma = \left(-\left.\frac{\mathrm{d}^2L}{\mathrm{d}X^2}\right|_{X_0}\right)^{-1/2}$$

The Gaussian, or *normal*, distribution



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

FWHM $\approx 2.35\,\sigma$

$$\langle x \rangle = \int x\,\mathrm{prob}(x|\mu,\sigma)\,\mathrm{d}x = \mu \quad \text{(mean)}$$

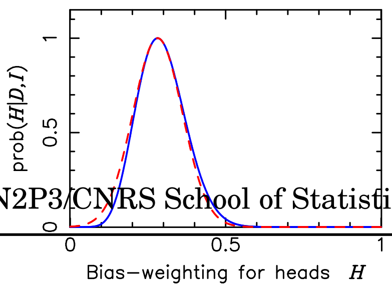
$$\left\langle (x-\mu)^2 \right\rangle = \int (x-\mu)^2\,\mathrm{prob}(x|\mu,\sigma)\,\mathrm{d}x = \sigma^2 \quad \text{(variance)}$$

Gaussian approximation for the coin

$\mathrm{prob}(H|\mathbf{D},I) \propto H^r(1-H)^{n-r} \;\Rightarrow\; L = \mathrm{const} + r\ln(H) + (n-r)\ln(1-H)$

$$\left.\frac{\mathrm{d}L}{\mathrm{d}H}\right|_{H_O} = \frac{r}{H_O} - \frac{(n-r)}{(1-H_O)} = 0 \quad \Rightarrow H_O = \frac{r}{n}$$

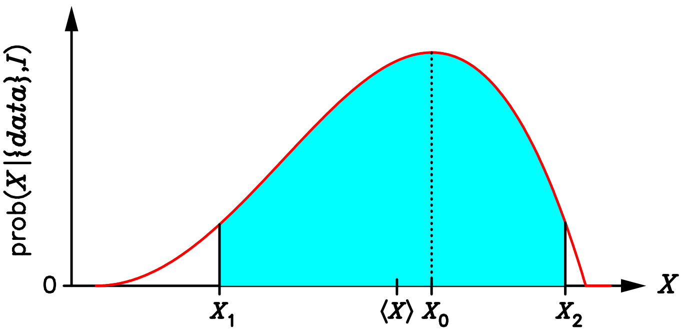
$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}H^2}\right|_{H_O} = -\frac{r}{H_O{}^2} - \frac{(n-r)}{(1-H_O)^2} = -\frac{n}{H_O(1-H_O)} \quad \Rightarrow \sigma = \sqrt{\frac{H_O(1-H_O)}{n}}$$



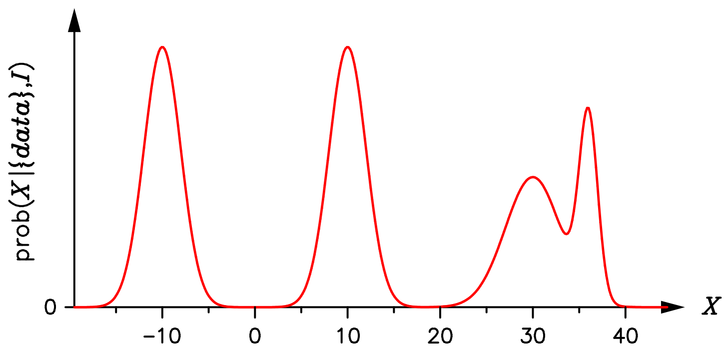
For $r=9$ and $n=32$,

$H \approx 0.28 \pm 0.08$

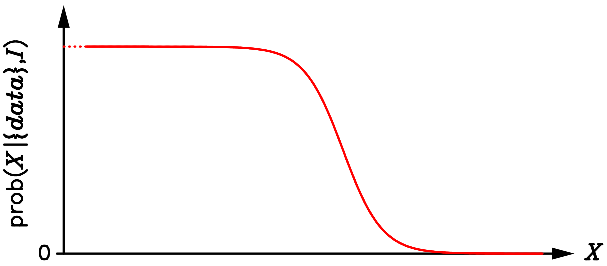
Asymmetric posterior pdfs



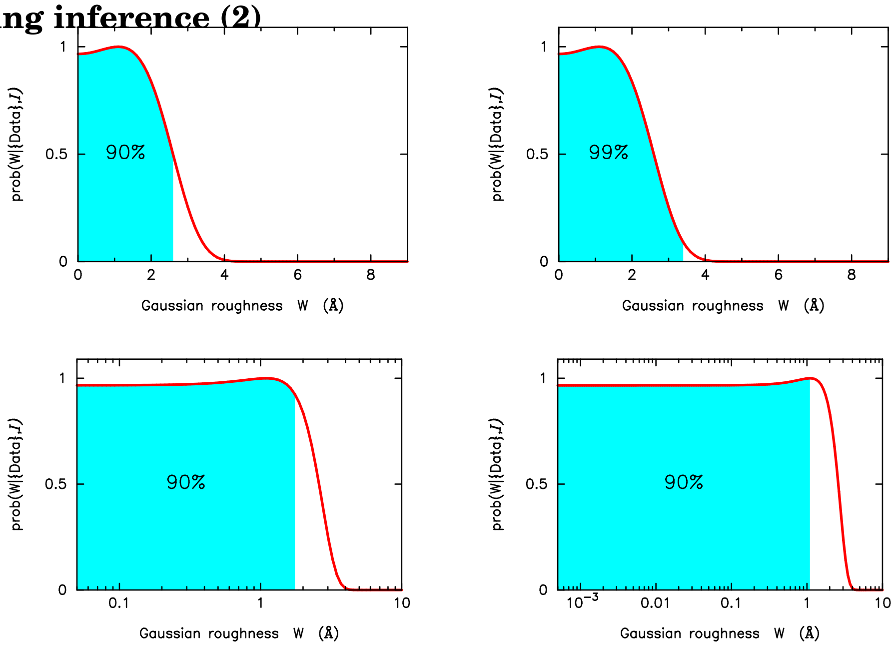
Multimodal posterior pdfs



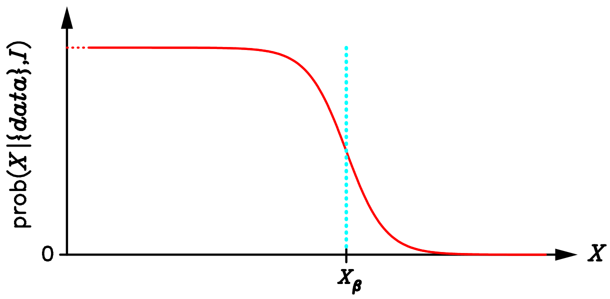
Limit-setting inference (1)



Limit-setting inference (2)



Limit-setting inference (3)



$$\frac{\mathrm{prob}(X\,|\,\{data\},I)}{[\mathrm{prob}(X\,|\,\{data\},I)]_{\max}} \begin{cases} > 1-\beta & \text{for } X < X_\beta \\ < 1-\beta & \text{for } X > X_\beta \end{cases}$$

$(0 < \beta < 1)$

Gaussian noise and averages (1)

Given N independent measurements of a quantity μ, $\{d_k\}$, all subject to Gaussian noise σ, what can we say about the value of μ?

- **Gaussian datum:** $\mathrm{prob}(d_k\,|\,\mu,\sigma,I) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left[-\dfrac{(d_k-\mu)^2}{2\,\sigma^2}\right]$

- **Independence:** $\mathrm{prob}(\{d_k\}\,|\,\mu,\sigma,I) = \prod_{k=1}^{N}\mathrm{prob}(d_k\,|\,\mu,\sigma,I)$

$$\mathrm{prob}(X,Y\,|\,I) = \underbrace{\mathrm{prob}(X\,|\,Y,I)}_{\mathrm{prob}(X\,|\,I)} \times \mathrm{prob}(Y\,|\,I) = \mathrm{prob}(X\,|\,I) \times \mathrm{prob}(Y\,|\,I)$$

- **Prior:** $\mathrm{prob}(\mu\,|\,\sigma,I) = \text{constant} \quad (\mu_{\min} \leqslant \mu \leqslant \mu_{\max})$

Gaussian noise and averages (2)

- Bayes: $\text{prob}(\mu|\{d_k\},\sigma,I) \propto \text{prob}(\{d_k\}|\mu,\sigma,I) \times \text{prob}(\mu|\sigma,I)$

$$\Rightarrow \; L = \log_e\big[\,\text{prob}(\mu|\{d_k\},\sigma,I)\,\big] = \text{const} - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(d_k-\mu)^2$$

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\mu}\right|_{\mu_o} = \sum_{k=1}^{N}\frac{d_k-\mu_o}{\sigma^2} = 0 \quad \Rightarrow \; \mu_o = \frac{1}{N}\sum_{k=1}^{N} d_k$$

$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2}\right|_{\mu_o} = -\sum_{k=1}^{N}\frac{1}{\sigma^2} = -\frac{N}{\sigma^2} \quad \Rightarrow \; \mu = \mu_o \pm \frac{\sigma}{\sqrt{N}}$$

Data with different-sized error-bars

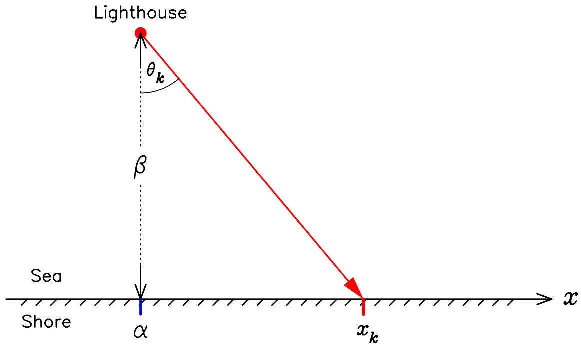
If the various measurements were of differing quality,
so that each datum d_k was associated with its own noise-level σ_k,
then the preceding analysis needs to be modified as follows.

$$L = \log_e\big[\,\text{prob}(\mu|\{d_k,\sigma_k\},I)\,\big] = \text{const} - \frac{1}{2}\sum_{k=1}^{N}\frac{(d_k-\mu)^2}{\sigma_k^2}$$

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\mu}\right|_{\mu_o} = 0 \quad \Rightarrow \; \mu_o = \sum_{k=1}^{N} w_k d_k \Bigg/ \sum_{k=1}^{N} w_k \,, \qquad \text{where } \; w_k = \frac{1}{\sigma_k^2}$$

$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2}\right|_{\mu_o} \Rightarrow \qquad \mu = \mu_o \pm \left(\sum_{k=1}^{N} w_k\right)^{-1/2}$$

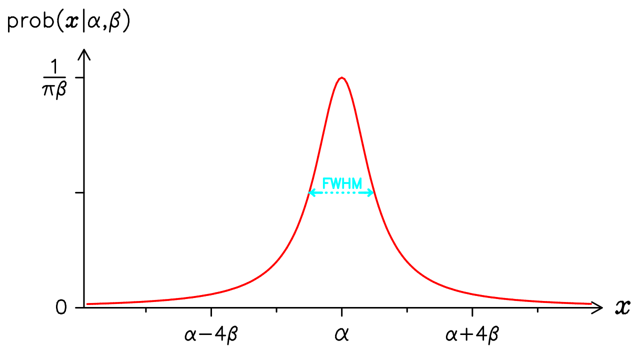
The lighthouse problem (1)



- **Data:** $\text{prob}(\theta_k|\alpha,\beta,I) = \frac{1}{\pi}$

But $\beta \tan\theta_k = x_k - \alpha \quad \Rightarrow \text{prob}(x_k|\alpha,\beta,I) = \dfrac{\beta}{\pi\left[\beta^2 + (x_k-\alpha)^2\right]}$

Lorentzian, or Cauchy, distribution



$$\text{prob}(x|\alpha,\beta) = \frac{\beta}{\pi\left[\beta^2 + (x-\alpha)^2\right]}$$

$$\text{FWHM} = 2\beta$$

$$\langle x \rangle = \int x \,\text{prob}(x|\alpha,\beta)\,\mathrm{d}x = \alpha \quad (*)$$

$$\left\langle (x-\alpha)^2 \right\rangle = \int (x-\alpha)^2 \,\text{prob}(x|\alpha,\beta)\,\mathrm{d}x \to \infty \quad (!)$$

The lighthouse problem (2)

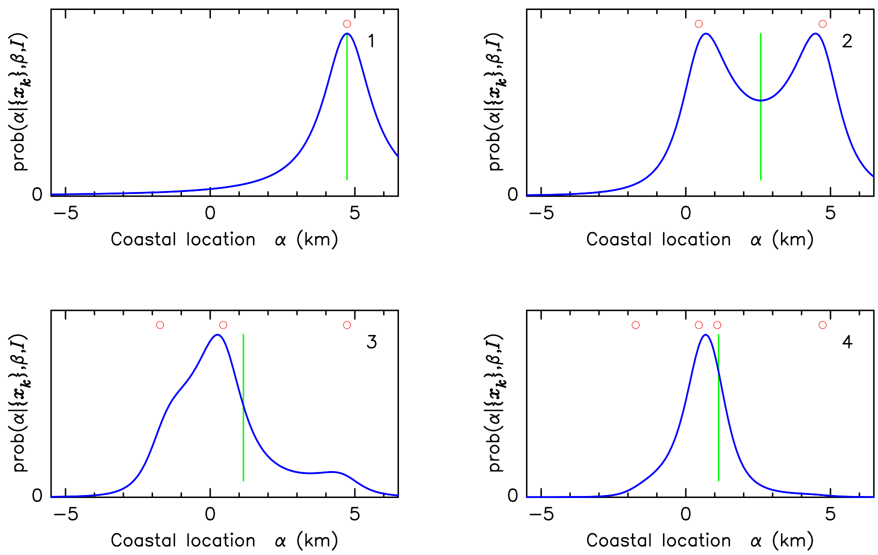
- Independence: $\text{prob}(\{x_k\}|\alpha,\beta,I) = \prod_{k=1}^{N} \text{prob}(x_k|\alpha,\beta,I)$
- Prior: $\text{prob}(\alpha|\beta,I) = \text{constant} \quad (\alpha_{\min} \leqslant \alpha \leqslant \alpha_{\max})$
- Bayes: $\text{prob}(\alpha|\{x_k\},\beta,I) \propto \text{prob}(\{x_k\}|\alpha,\beta,I) \times \text{prob}(\alpha|\beta,I)$

$$\Rightarrow L = \log_e\left[\text{prob}(\alpha|\{x_k\},\sigma,I)\right] = \text{const} - \sum_{k=1}^{N} \log_e\left[\beta^2 + (x_k-\alpha)^2\right]$$

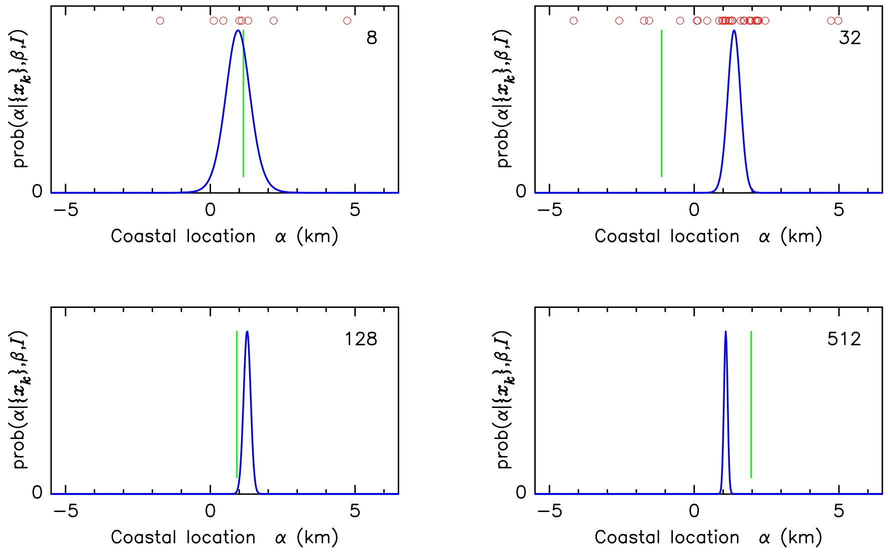
$$\left.\frac{\mathrm{d}L}{\mathrm{d}\alpha}\right|_{\alpha_o} = 2\sum_{k=1}^{N} \frac{x_k-\alpha_o}{\beta^2+(x_k-\alpha_o)^2} = 0$$

Can't be solved analytically!

Lighthouse example (1)



Lighthouse example (2)



Moral of the lighthouse story

The sample *mean* is **not** always a useful number.

(Infinite variance of Cauchy ⇒ *central limit theorem* does not hold)

Let probability theory decide what's best!