

# Occam's Razor and Bayesian Analysis

Bill Barnard – May 17, 2005

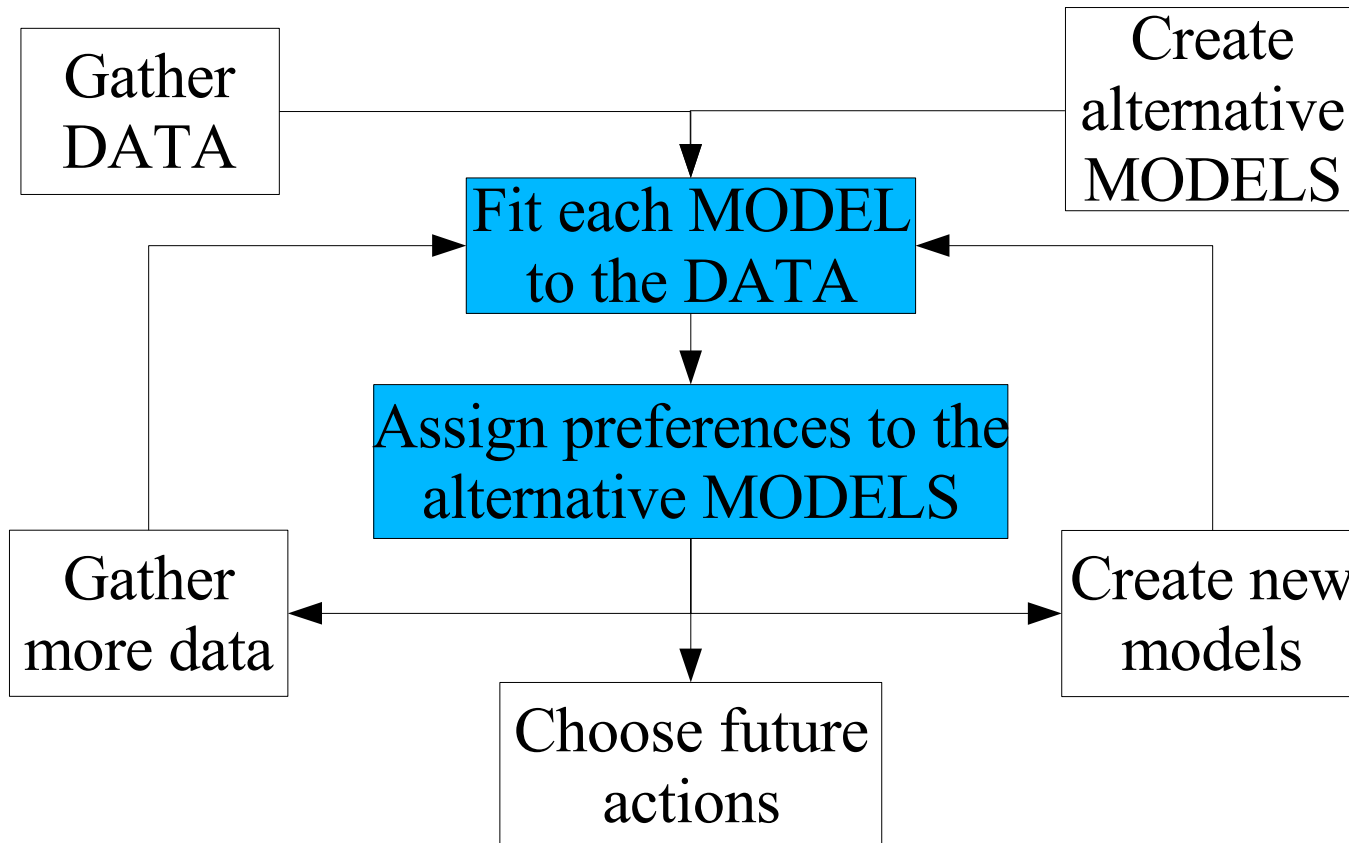
# Probability: frequency vs. plausibility

- Probability theory originated from mathematics of gambling.  
Probability = frequency of outcome in a series of identical trials.
- Bayesian probability is the measure of the plausibility of a hypothesis.
  - Dinosaurs *probably* died out due to climatic change.
  - Rain is expected with a 30% *probability*.
- Each way of looking at the world has merit, and consequent areas of application.

# Induction vs. Deduction

- Deduction – reasoning in the presence of certainty.
  - if X implies Y, and X is true, then Y must be true.
- Induction – reasoning in the presence of uncertainty.
  - if X implies Y, and Y is true, then X is *more plausible*.
- The Black Swan Problem - “*No amount of observations of white swans can allow the inference that all swans are white, but the observation of a single black swan is sufficient to refute that conclusion.*”
- Theories may be known to be wrong (falsified), or not falsified but subject to test for falsification.

# Bayesian Inference in Data Modeling



# Bayesian Inference in Data Modeling

- 1<sup>st</sup> level: Model fitting, e.g. parameter estimation
  - formal use of prior information.
  - often differs little from a 'frequentist' approach.
  - maximum likelihood methods can lead to overfitting.
- 2<sup>nd</sup> level: Model comparison
  - ranks models objectively, based on the observed data.
  - usually does not use prior information.
  - automatically embodies Occam's razor.

# Model Fitting

*Posterior Probability Distribution*

$$P(\theta|D, M) = \frac{P(D|\theta, M) \cdot P(\theta)}{P(D|M)} \quad \text{where} \quad P(D|M) = \int_{\theta'} P(D|\theta') \cdot P(\theta') \cdot d\theta'$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

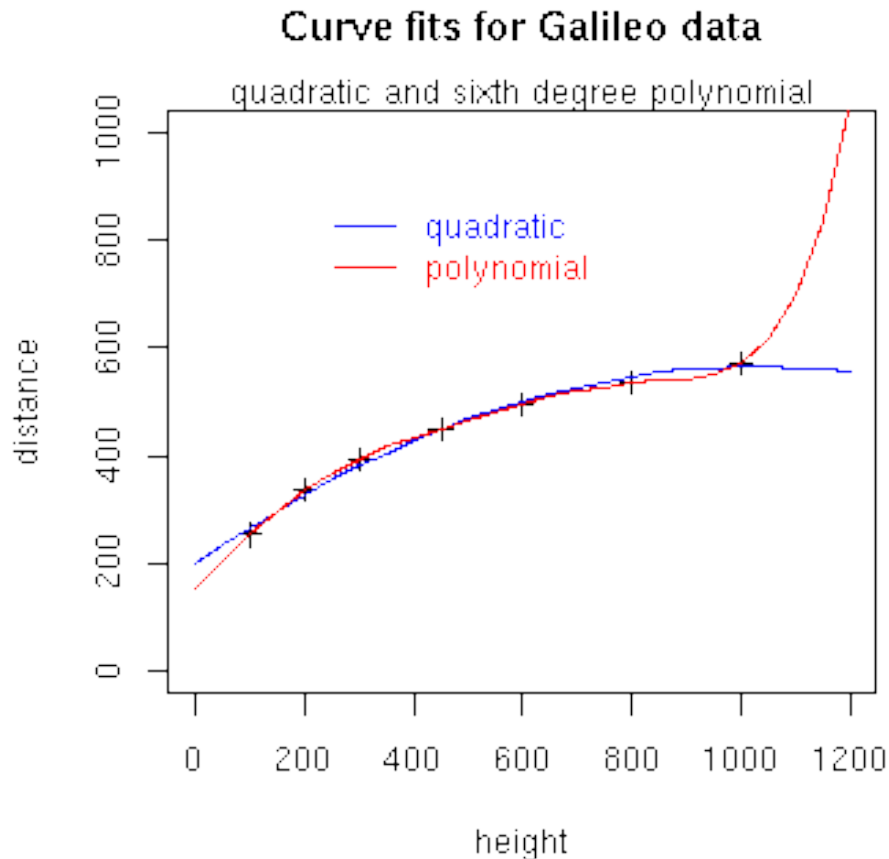
*Maximum Likelihood (ML) Estimation*  $\theta^{ML} = \underset{\theta}{\operatorname{argmax}} P(D|\theta, M)$

*Maximum a Posteriori (MAP) Estimation*  $\theta^{MAP} = \underset{\theta}{\operatorname{argmax}} P(D|\theta, M) \cdot P(\theta|M)$

- ML is a frequentist method. Appears to be used more often than MAP. Danger of overfitting unless sufficient data are available.
- MAP often differs little from ML, especially if 'flat' priors are chosen. No general, objective principle has been accepted for assigning prior probabilities.

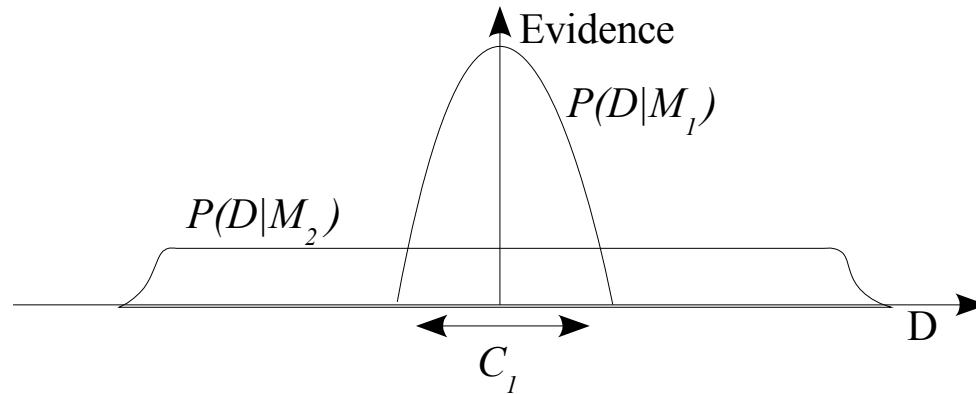
# Model Fitting

## example



- Data collected by Galileo in 1608 – ball rolling down an inclined plane, then continuing in free-fall.
- Occam's razor suggests the simpler model is better; it has a higher prior probability.
- The simpler model may have a greater posterior probability (the plausibility of the model)

# Why Bayes embodies Occam's razor



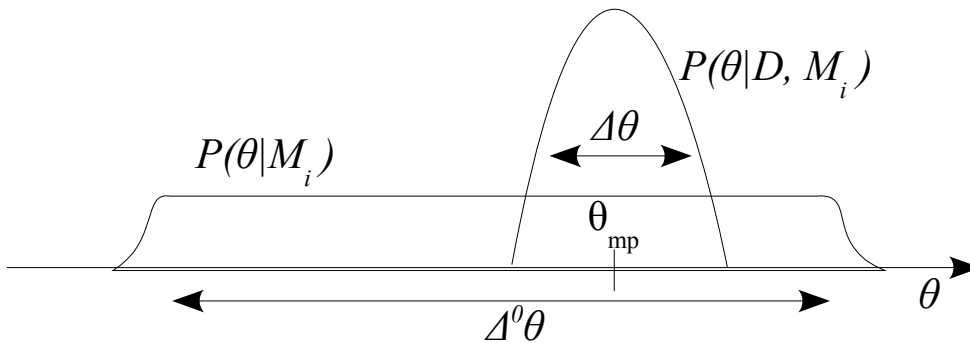
- $P(D|M_i)$  is known as the **evidence** for  $M_i$ .
- Complex models are penalized. Bayes rewards models according to how well they predict actual data.
- The simpler, less powerful model  $M_1$  makes a sharper prediction. If data fall in region  $C_1$ , then  $M_1$  is more probable than  $M_2$ .

# Model Comparison

*Posterior probability of model  $M_i$ :  $P(M_i|D) \propto P(D|M_i) \cdot P(M_i)$*

- Probability of a model is a measure of the model's plausibility and not a measure of frequency of occurrence.
- Assign preferences by evaluating the evidence  $P(D|M_i)$ .
  - parametric or non-parametric
- Unequal priors,  $P(M_i)$  may be tried; there is no “right” prior.
  - the “sure thing” hypothesis has a tiny prior probability.
  - prior probabilities are selected before examining data
  - as quantity of data becomes large,  $P(D|M_i)$  will always be more important than  $P(M_i)$ .

# Evaluating the evidence



The prior distribution has width  $\Delta^0\theta$ .  
The posterior distribution has a single peak at  $\theta_{mp}$  with width  $\Delta\theta$ .

The Occam factor is  $\Delta\theta / \Delta^0\theta$

*Evidence is the normalizing constant from the posterior probability for :  $\theta$*

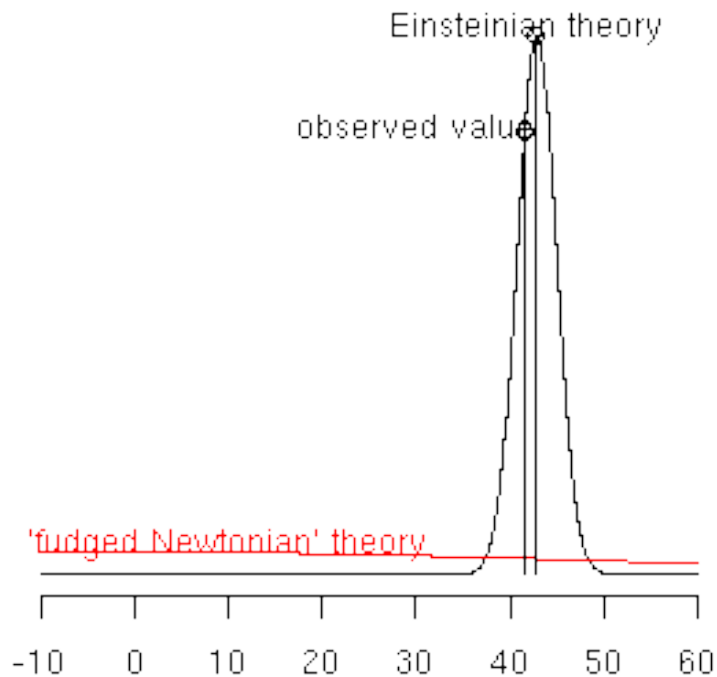
$$P(D|M_i) = \int P(D|\theta, M_i) P(\theta|M_i) \cdot d\theta$$

$$\underbrace{P(D|M_i)}_{\text{Evidence}} \simeq \underbrace{P(D|\theta_{mp}, M_i)}_{\text{Best fit likelihood}} \underbrace{P(\theta_{mp}|M_i) \Delta\theta}_{\text{Occam factor}}$$

*Evidence is found by taking the best fit likelihood the model can achieve and multiplying it by an 'Occam factor', a term with magnitude less than one that penalizes  $M_i$  for having the parameter  $\theta$ .*

# Model Comparison example

## Mercury's anomalous motion



anomalous perihelion motion (arc-seconds per century)

Anomalous perihelion advance observed in the motion of Mercury can be explained by two hypotheses. Einstein's general theory of relativity makes a sharp prediction of 42.9 arc-seconds per century. The “fudged Newtonian” theory can be adjusted to accommodate almost any observation.

Einstein curve is  $N(\mu = 42.9, \sigma = 2.07)$  where  $\sigma$  is determined from the probable error of the observed value  $a$  ( $41.6 \pm 1.4$ ). Probable error =  $0.6745\sigma$ .  
 $P(a | E) = 0.158$

“Fudged Newtonian” curve is  $N(\mu = 0, \sigma = 50.04)$ . This was determined by a more involved analysis (Jefferys & Berger).  
 $P(a | F) = 0.00564$ .

The Einsteinian theory is favored by odds of 28.0:1. Worst case bounds favor Einstein by 27.76:1 or 15:1.

# Conclusions

- Comparison of evidence,  $P(D|M)$ , provides a purely objective way to rank hypotheses.
- Evaluation of evidence is an extension of maximum likelihood model selection: multiply the best fit likelihood by the Occam factor. No more computationally difficult than finding the best fit parameters.
- The Occam factor automatically penalizes a model which requires fine tuning of its parameters. It promotes models where the required precision of its parameters is coarse.

# References

- Baldi, P. and Brunak, S. 2001. *Bioinformatics – The Machine Learning Approach*. 2<sup>nd</sup> ed. MIT Press.
- Durbin. R., Eddy, S., Krogh, A., Mitchison, G. 1998. - *Biological Sequence Analysis*. Cambridge University Press.
- Jefferys, W.H. and Berger, J.O. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64-72.
- MacKay, D.J.C. 1992. Bayesian Interpolation. *Neural Computation* 4:415-447.  
see also <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>
- Taleb, N.N. 2004. *Fooled by Randomness – The Hidden Role of Chance in Life and in the Markets*. 2<sup>nd</sup> ed. Thomson-Texere.