

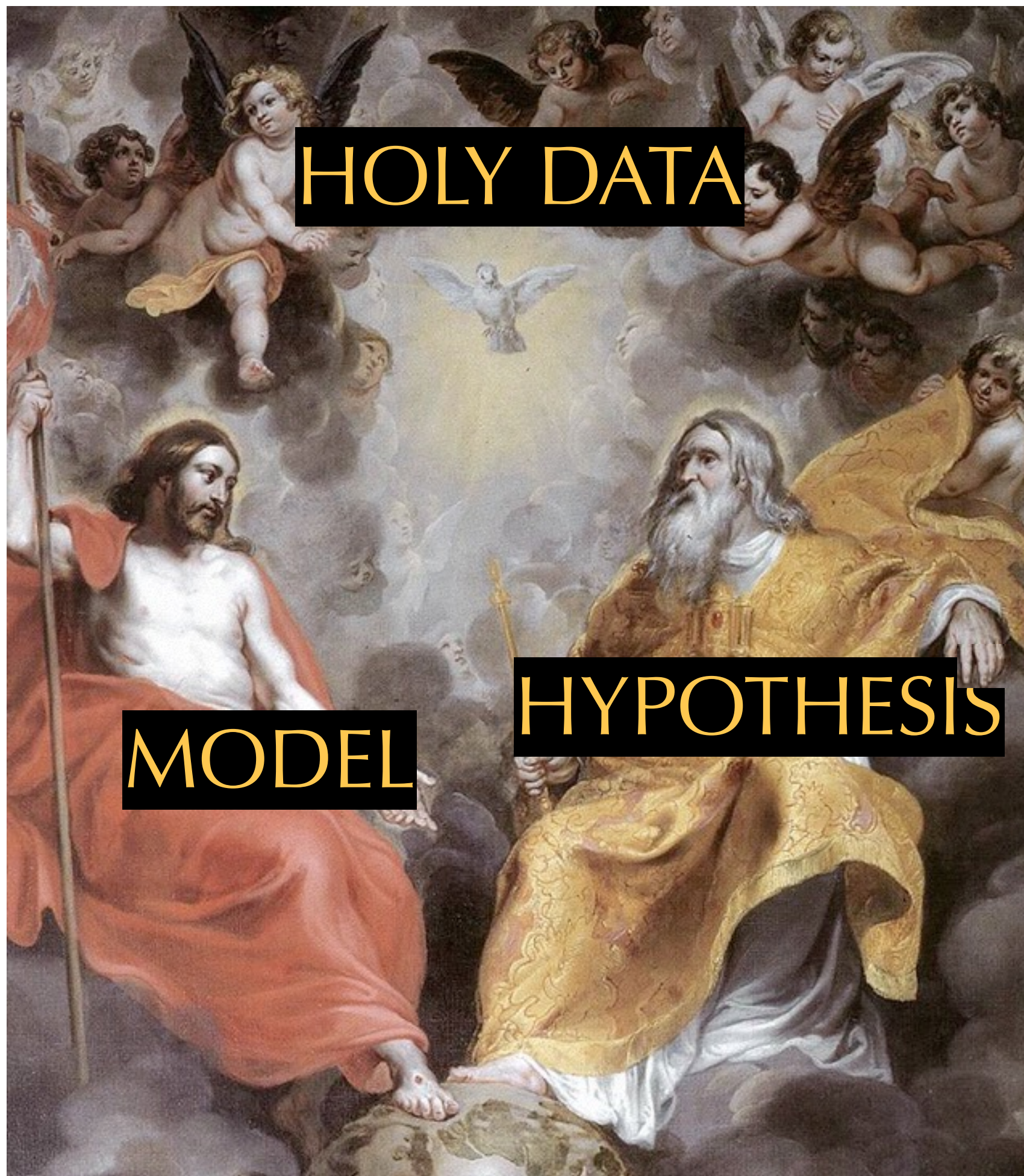
Basic principles of statistical inference

Oliver Pybus

What Is Statistical Inference?

“The use of a sample of data to draw inferences or conclusions about some aspect of the situation from which the data were taken.”





*Holy Trinity
Hendrik van
Balen, 1620*


Data

- Typically an alignment of gene sequences, with date and location of sampling for each. Sometimes phenotypic trait data is available.
- Population genetic inference usually requires that sequences are approximately randomly sampled.
- Phylogenetic inference problems often don't require random sampling.
- The data is assumed to be correct, although uncertainty in the data can often be modelled (e.g. ambiguous nucleotides).
- Alignment uncertainty is usually ignored.

Model

- A mathematical representation of the situation under study (or of some aspect of the situation).
- There are several types of evolutionary models, often used in combination.
 - *Nucleotide substitution models* (e.g. JC, HKY, GTR)
 - *Molecular clock models* (e.g. strict, relaxed, local)
 - *Trait evolution models* (e.g. strict, relaxed random walks)
 - *Population models* (e.g. coalescent, BD models)
- Each model has a number of model **parameters**.
- Is a tree topology a model, or a set of parameters?

Hypothesis

- A theory about the situation under study.
- Mathematically, a hypothesis is some statement about the parameter values of the model.
 - e.g. “the dN/dS of my sequences is > 1 ”
 - e.g. “the date of this phylogenetic node is 1982”
 - e.g. “the tree topology of my sequences is  ”
- The data can be used to assess whether a given hypothesis is reasonable or not.
- The model may have many parameters, but the hypothesis may not concern all of them. The remainder are called **nuisance parameters**.

Goals of Inference

POINT ESTIMATION

- Using the data to estimate values for one or more model parameters.
- e.g. “When was the most recent common ancestor of my sampled sequences?”
 - *Data*: sequence alignment, dates of sampling, tree topology
 - *Model*: HKY
 - *Estimated parameter*: age of root node
 - *Nuisance parameters*: ages of the other nodes in the tree

Goals of Inference

INTERVAL ESTIMATION

- Using the data to provide a range that represents the degree of uncertainty in the estimate of a parameter.

e.g. “What are the 95% confidence intervals for the evolutionary rate of my sequences?”

Goals of Inference

HYPOTHESIS TESTING

- Using the data to measure the relative plausibility of different statements about the model parameters.

e.g. “Is my estimated dN/dS value significantly greater than 1?”

e.g. “Does REV fit my data better than HKY?”

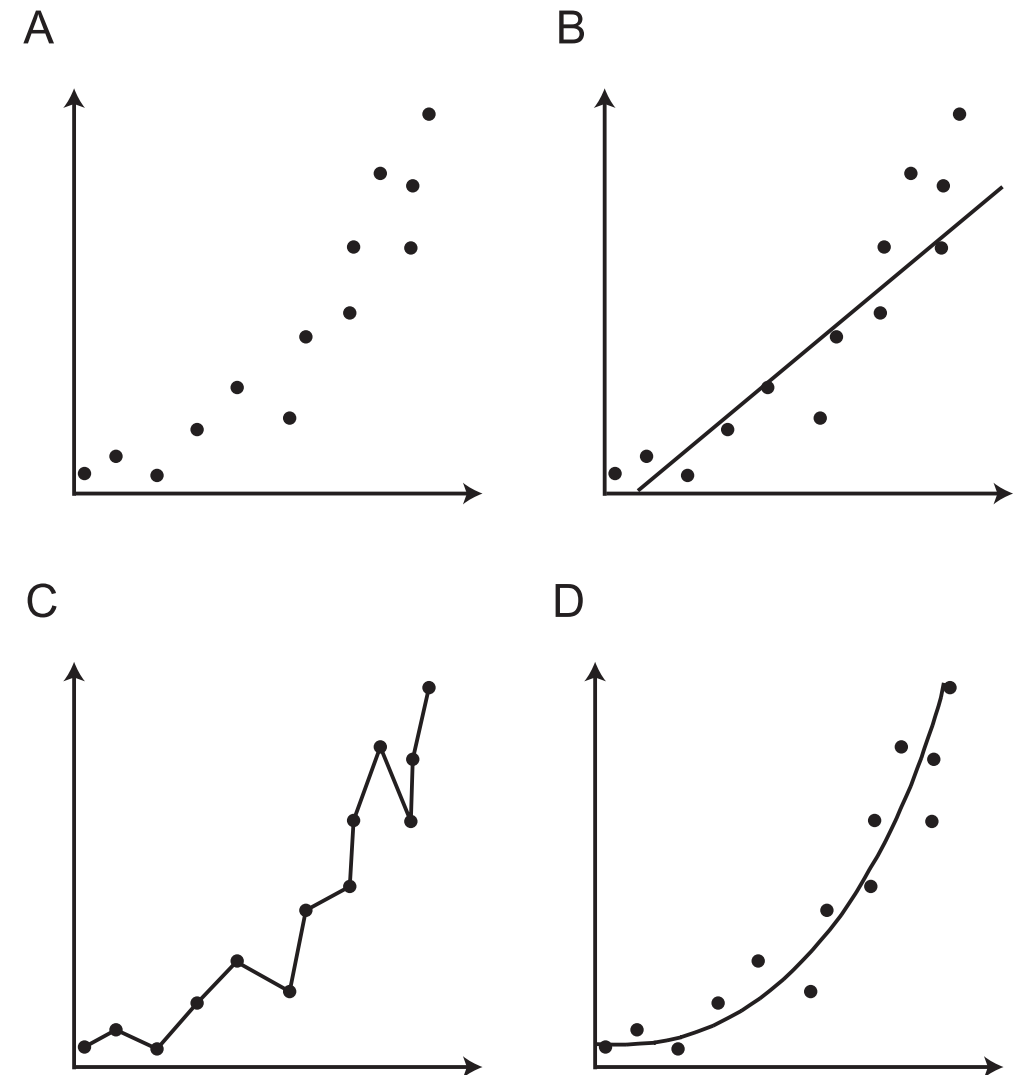
e.g. “Can I reject a strict molecular clock?”

e.g. “Do sequences A, B & C form a clade in my tree?”

Goals of Inference

MODEL SELECTION

- The process of finding the most appropriate model for your data.
- It involves a trade-off between “goodness of fit” and “predictive power”.
- Adding parameters increases the former but decreases the latter.
- Remember, even the best-fitting model may explain the data poorly.



Inference Frameworks

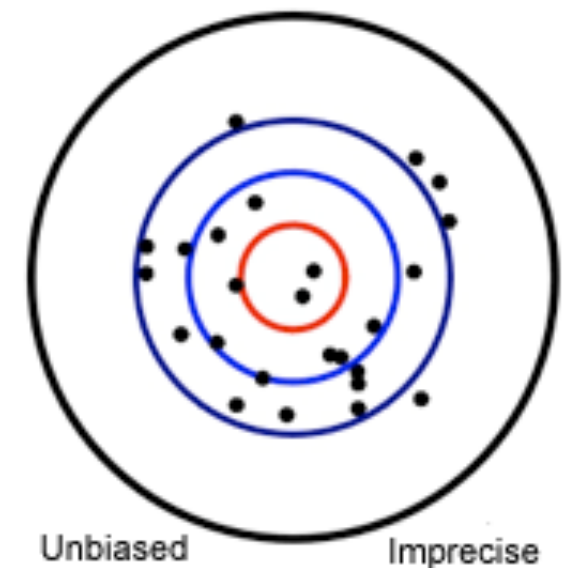
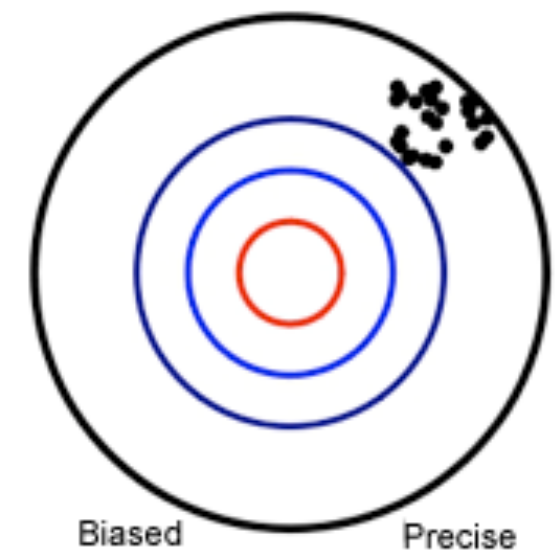
- There are several different types of statistical inference.
- Evolutionary problems often involve complex models with many parameters, and sometimes limited amounts of data.
- **Likelihood** and **Bayesian** inference are practical methods in this situation. They are related, but differ philosophically in the way they view probability.

Inference Frameworks

- Likelihood (frequentist) inference: Probabilities refer only to the outcome of experiments (i.e. data). They represent the frequencies of outcomes if the experiment were repeated many times. The degree to which data supports a hypothesis is termed *likelihood*.
- Bayesian inference: Both data and model parameters are described by probabilities. Probability reflects our degree of belief in a hypothesis, as well as representing the outcome of experiments. Hence hypotheses have probabilities even in the absence of data.

Properties of Inference Methods

- *BIAS*: The average deviation of an estimate from the true value.
- *VARIANCE*: Imprecision, or the degree of uncertainty in an estimate. Reflected in large confidence intervals, or in a wide “spread” of values when estimation is re-run many times.
- *CONSISTENCY*: The convergence of an estimate to the true parameter value as sample size increases.
- *ERROR*: The failure of hypothesis tests to get the right answer as often as they should.



Likelihood

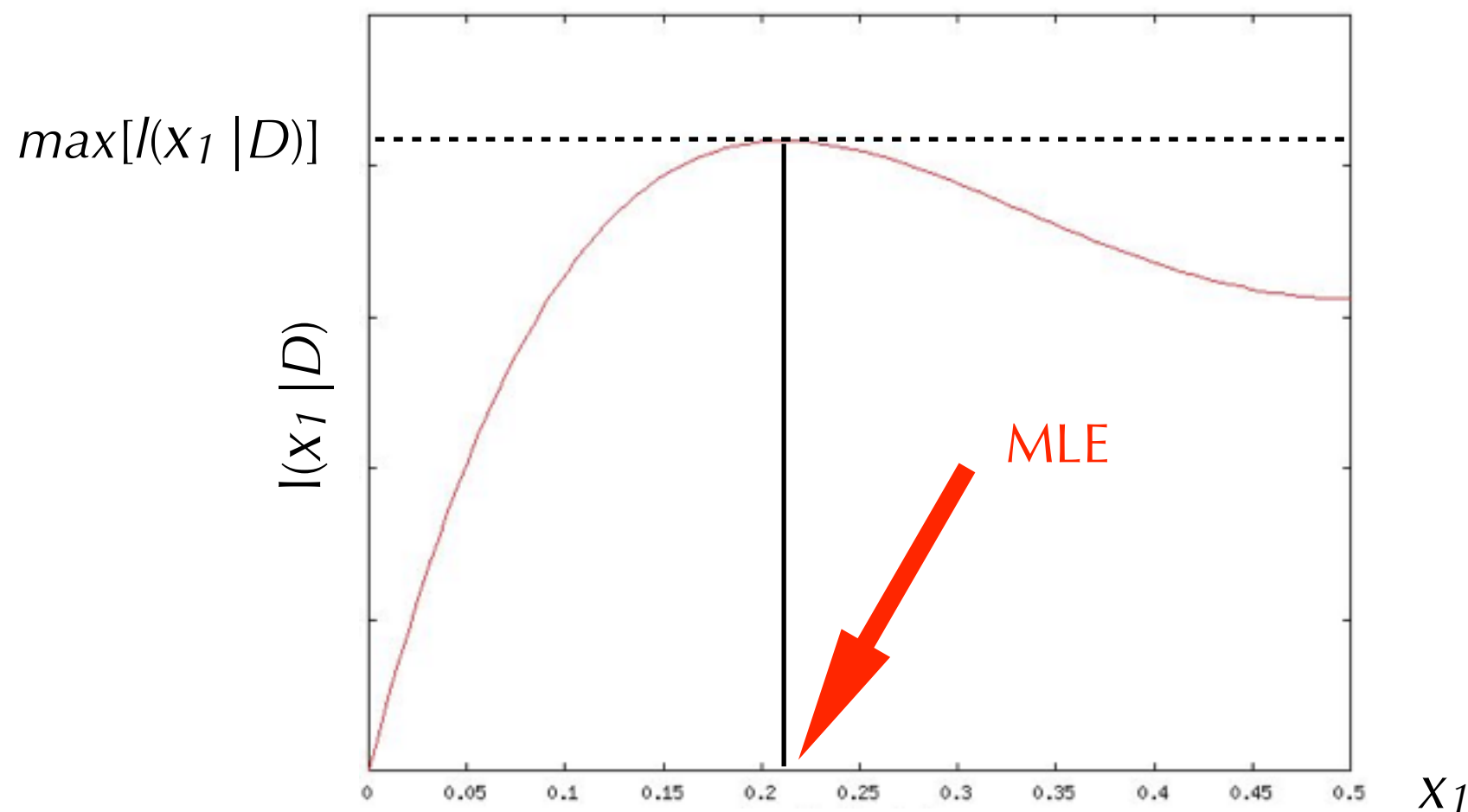
- “A technique for assessing the relative merits of different hypotheses in the light of the data.”
- The probability of observing data D , given hypothesis H , is denoted $P(D|H)$. It is defined by the model and is a probability density function.
- The likelihood of hypothesis H , given data D , is denoted $L(H|D)$. This is a likelihood function and is directly proportional to $P(D|H)$.
i.e. $L(H|D) \propto P(D|H)$
- The constant of proportionality is arbitrary, but is the same for all hypotheses under the same data.
- Unlike probabilities, likelihoods do not sum to 1!

Likelihood Ratios

- Inferences are made by comparing the likelihoods of different hypotheses on the same data (D).
e.g. hypotheses H_1 & H_2 have likelihoods $L(H_1|D)$ & $L(H_2|D)$
- **NEVER COMPARE LIKELIHOODS ON DIFFERENT DATA!**
- Likelihoods are very small numbers, so the natural logarithm of the likelihood, denoted $l(H|D)$, is used.
- Likelihood ratios are therefore log-likelihood differences.
i.e. $L(H_1|D) / L(H_2|D) = l(H_1|D) - l(H_2|D)$

Maximum Likelihood Estimation

- Suppose we have a model with one parameter, x_I .
- We find the value of x_I that maximises the likelihood. This value is the maximum likelihood estimate (MLE) of x_I .
- A plot of x_I against log-likelihood is a log-likelihood curve.
- Algebra or optimisation algorithms find the highest point of the curve.



Maximum Likelihood Estimation

Situation:

Flipping a coin $n=10$ times

Data:

$h=6$ Heads, $(n-h)=4$ Tails

Model:

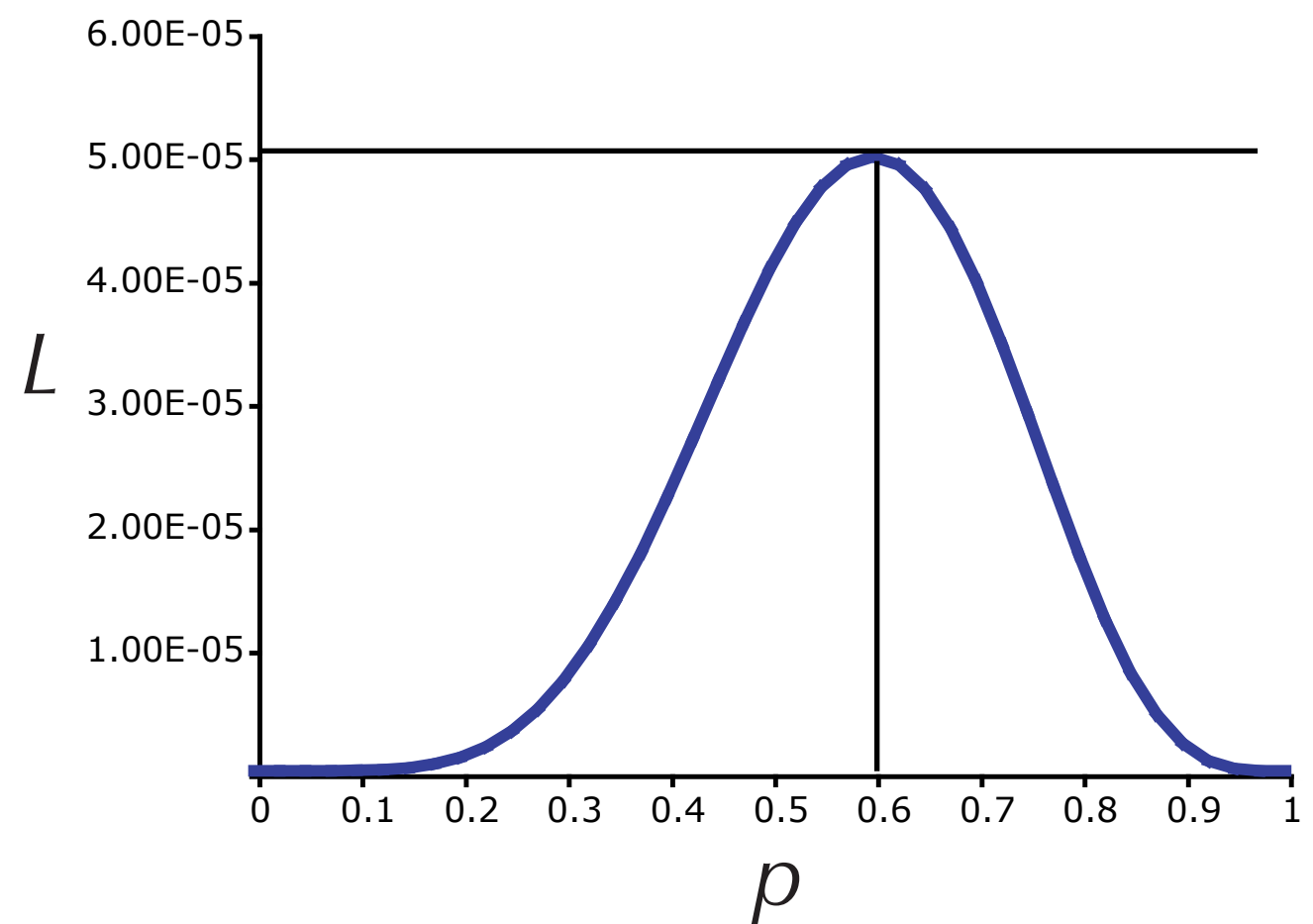
binomial distribution

Likelihood function:

$$L[p|h,n] = \binom{n}{h} p^h (1-p)^{n-h}$$

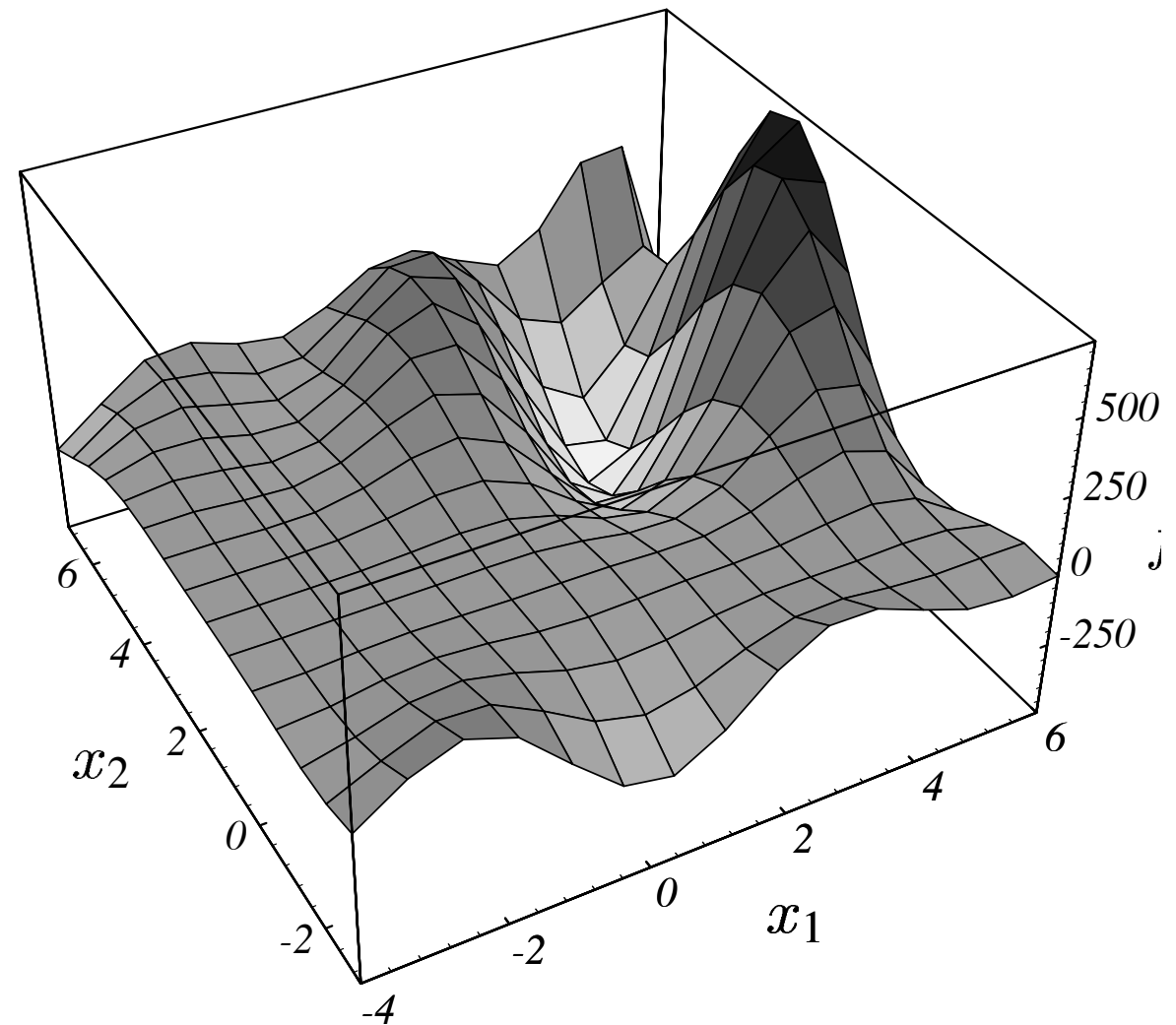
MLE:

$$\hat{p} = h/n$$



Maximum Likelihood Estimation

- If we have two parameters, x_1 & x_2 , then the pair of x_1 & x_2 values that maximises the likelihood function must be found.
- Optimisation algorithms try to find the maximum point.
- This can be a computationally difficult problem.



Likelihood Ratio Statistic

- The likelihood ratio statistic (LRS) is used to compare hypotheses.
- Suppose we have two hypotheses, H_1 & H_2 . These can represent specific parameter values, or regions of parameter space, or whole models.
- $$\text{LRS} = 2 * \{ \max[l(H_1|D)] - \max[l(H_2|D)] \}$$

Likelihood Ratio Statistic

- If the hypotheses are nested (one is a special case of the other) then the LRS can be used to compare their goodness-of-fit.*
- The LRS is also used to calculate confidence limits for MLEs.*
- Non-nested hypotheses can be compared using the **Akaike Information Criterion (AIC)**. For hypothesis H :
$$AIC(H) = \max[l(H|D)] - n$$
where n is the number of parameters in the model.
- “Better” hypotheses have higher AIC values.

* Not covered here.

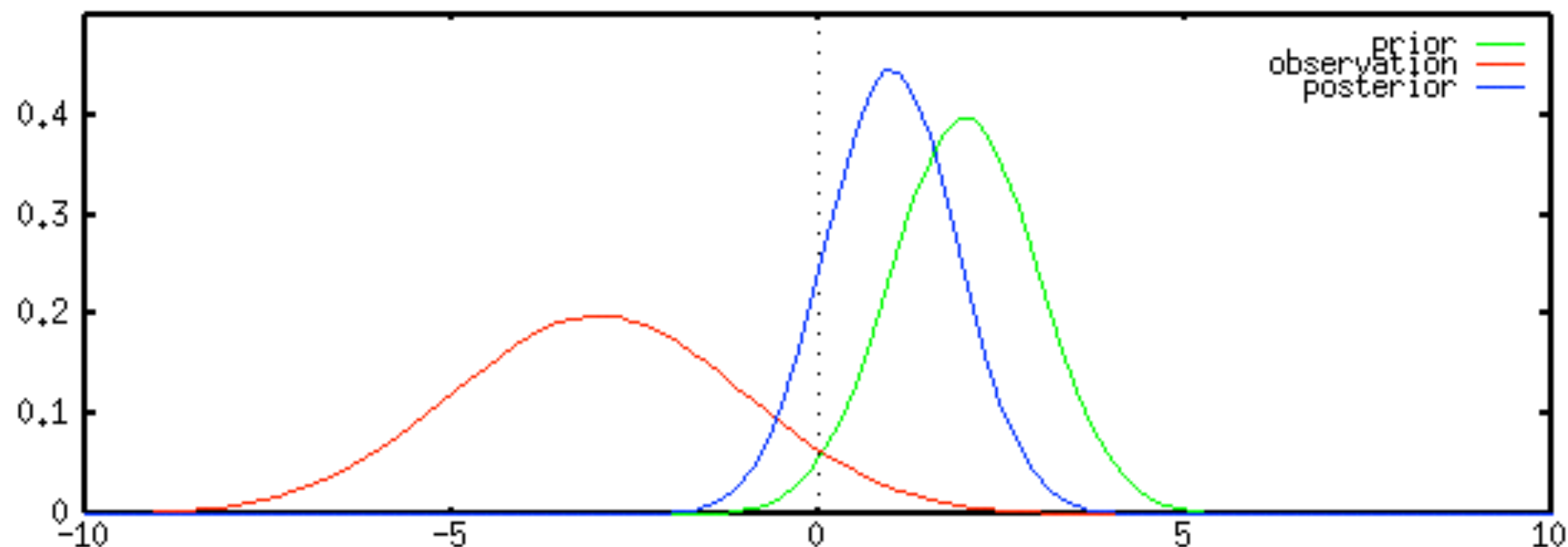
Bayesian Inference

- Bayesian inference produces a **posterior probability distribution** rather than a likelihood curve.
- The “posterior” combines information from the data *and* from previous knowledge. (Likelihood = the data only.)
- Each parameter in the model has a **prior probability distribution** representing our previous knowledge or belief about that parameter.
- The “prior” can be strong or weak...
 - e.g. “Human heights follow a normal distribution with mean = 1.7m and standard deviation = 15cm”
 - e.g. “Human heights follow a uniform (flat) distribution between 10^{-10} m and 10^{26} m.”

Bayesian Inference

If the posterior and the prior look similar then the data is not very informative about the parameter in question.

We don't directly compare posteriors and priors as often as we should.



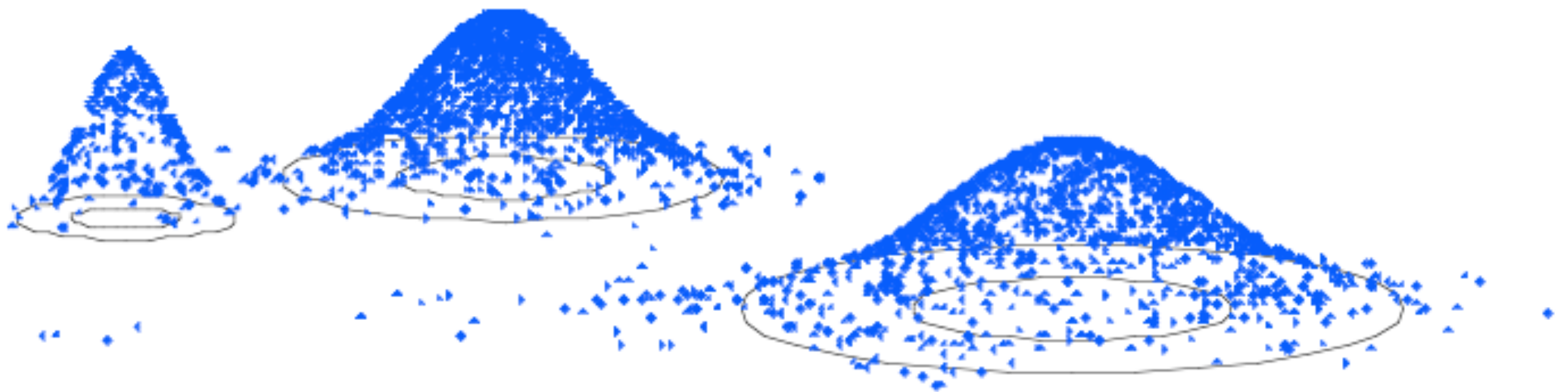
Posterior distributions are defined using **Bayes' Theorem**.*

* Not covered in this lecture.

Posterior Distributions

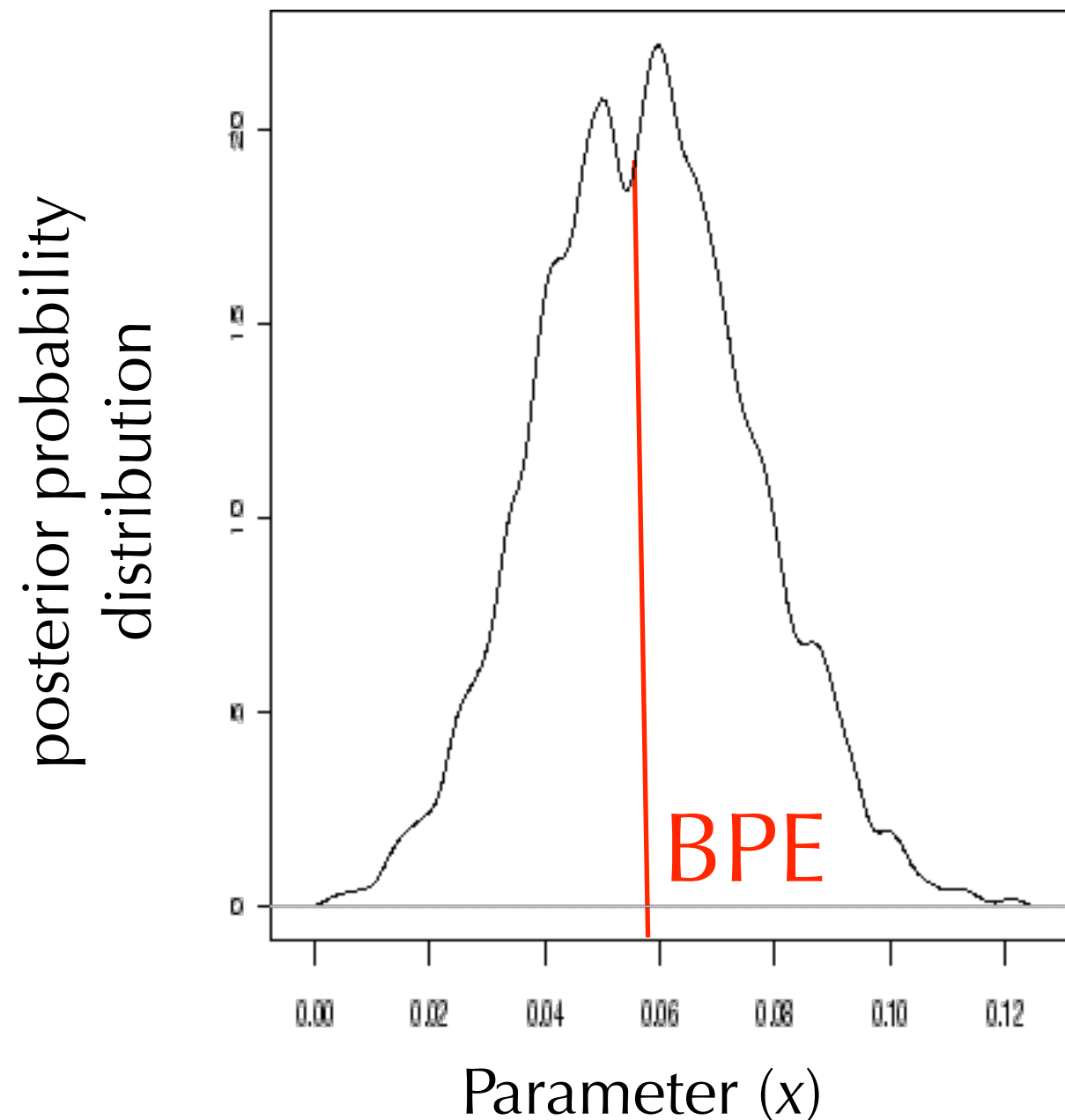
Posterior distributions are very difficult to calculate directly. However, we can approximate the posterior distribution by using **Markov Chain Monte Carlo (MCMC)** sampling.

This algorithm walks around parameter space in a pseudo-random way. It moves quickly through 'low' regions and slowly in 'high' regions whilst keeping a record of where it has been.



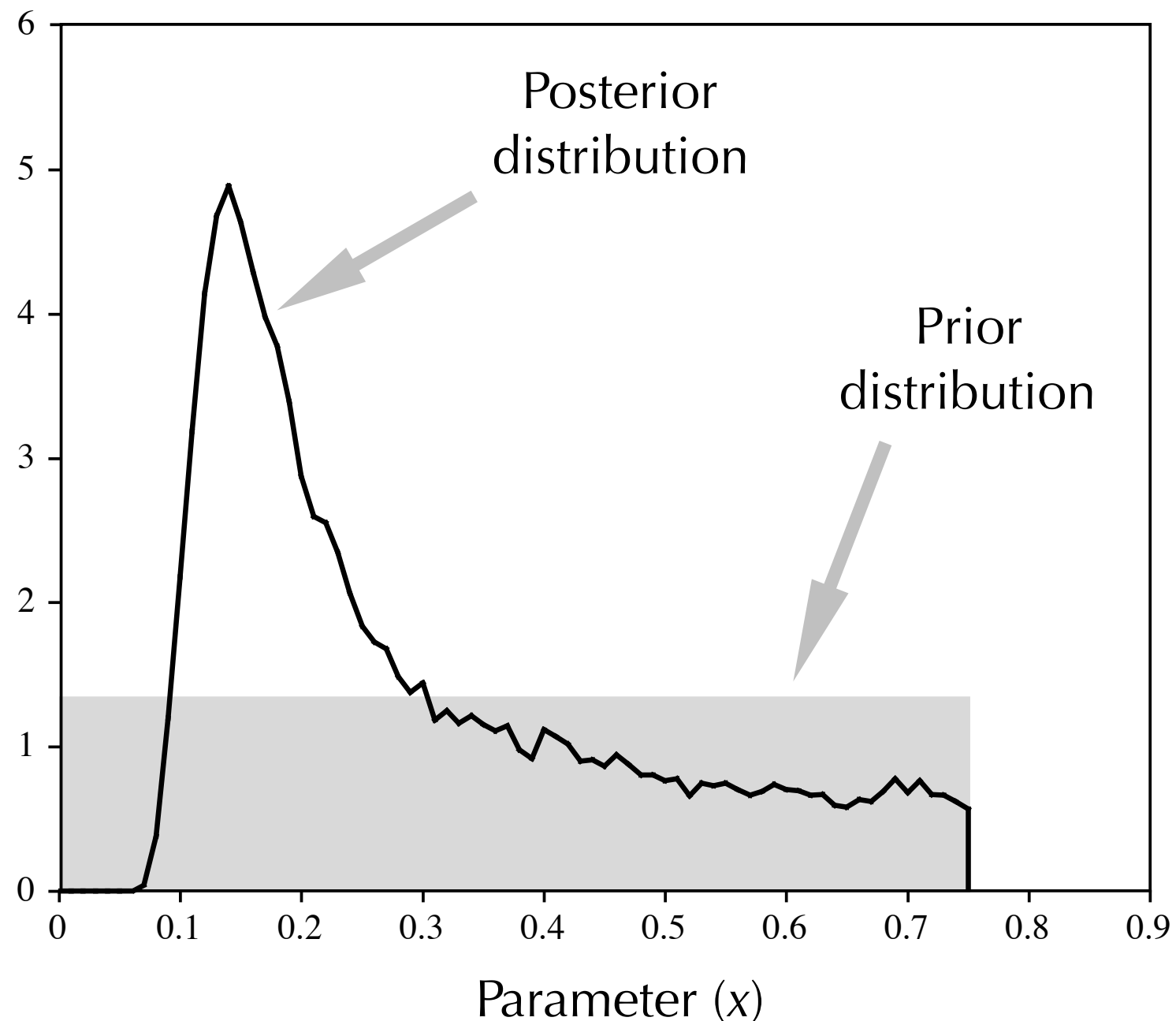
Posterior Distributions

Parameters are estimated by finding the mean or median of the posterior distribution. These are Bayesian posterior estimates (BPEs).



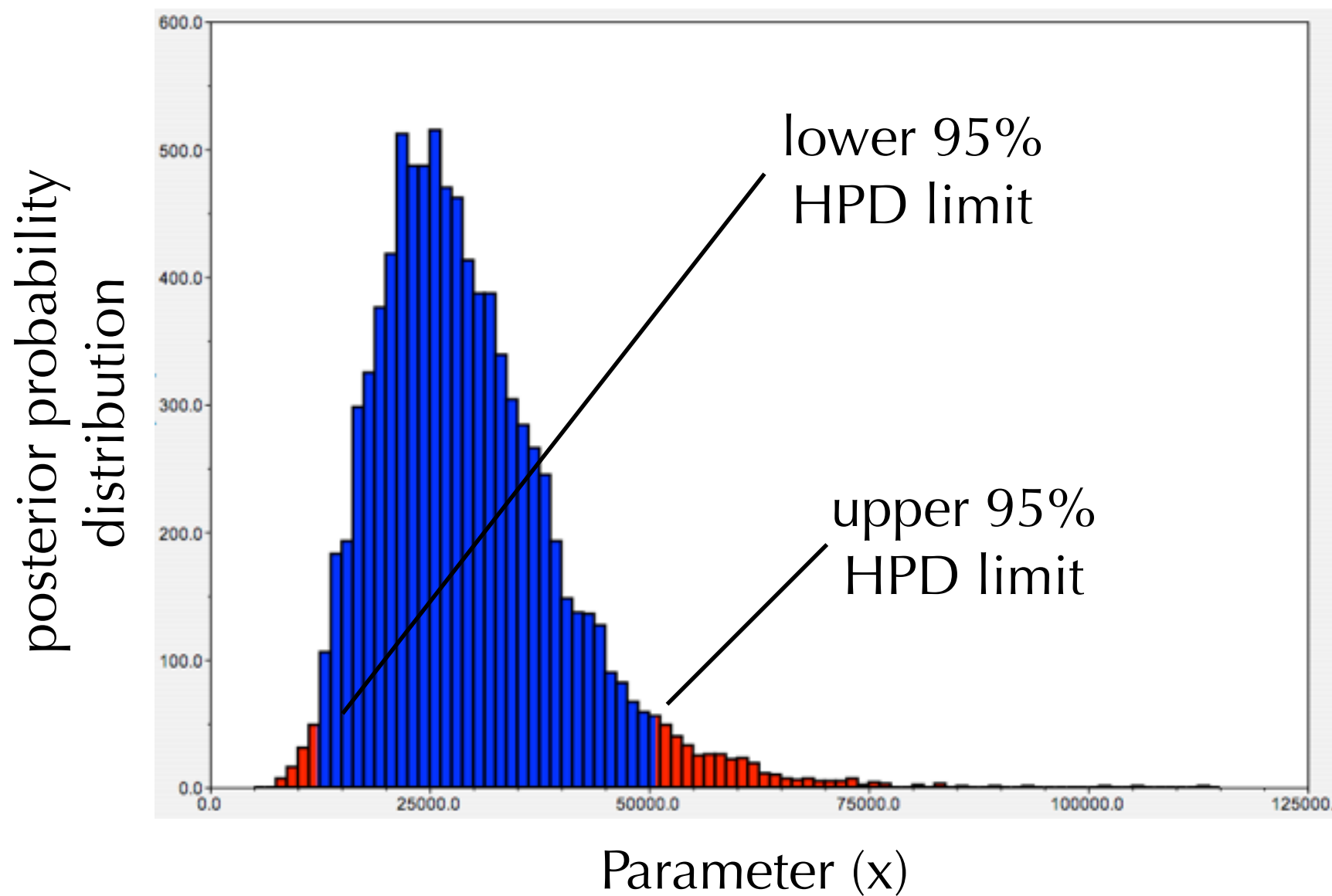
Posterior Distributions

Here, the posterior is cut off at $x=0.75$ by the prior. The data support values of $x>0.75$, but the prior won't allow this.

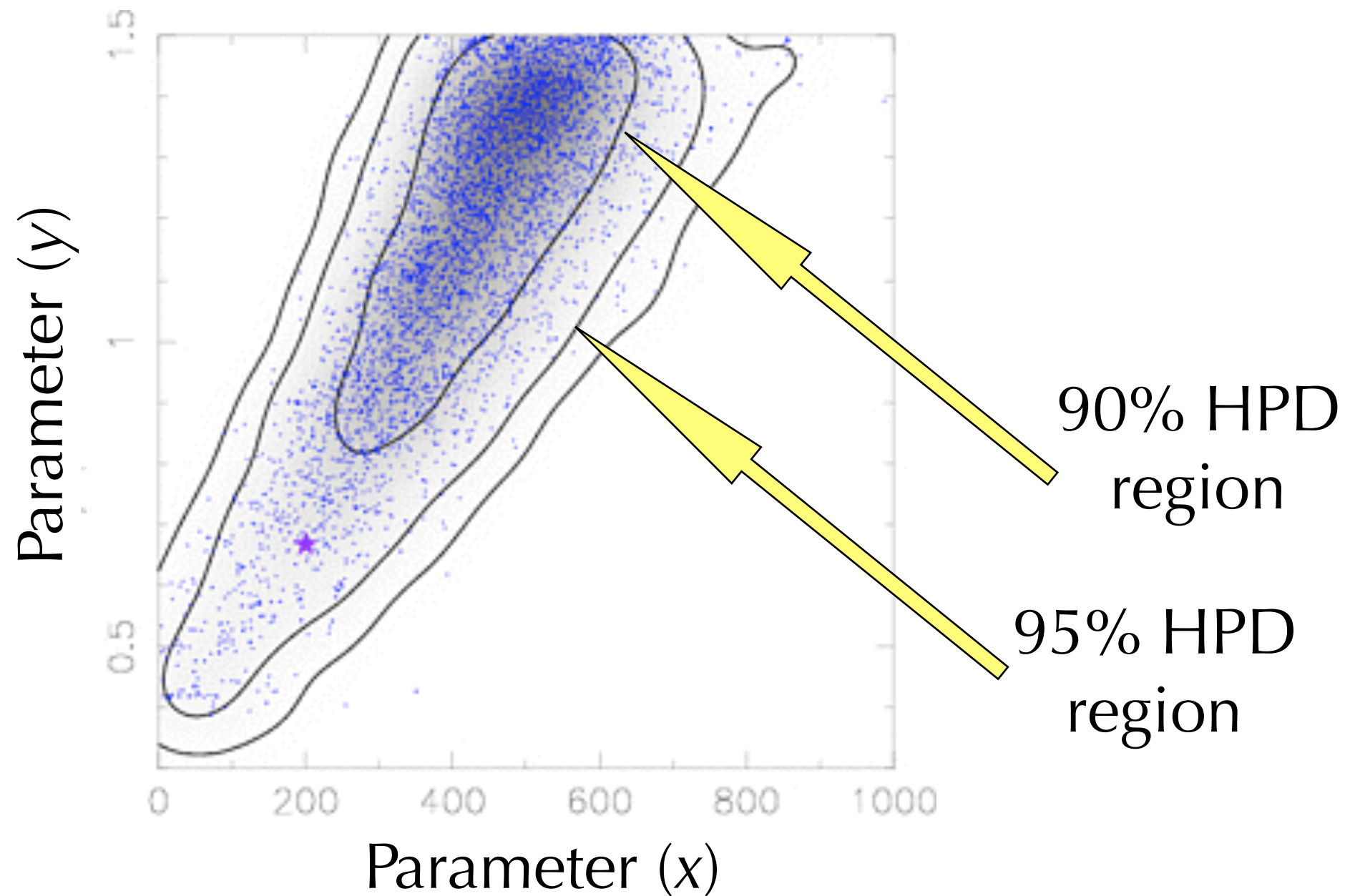


Credible Regions

The Bayesian equivalent of a confidence interval is called the highest posterior density (HPD) credible region. This is the smallest region that contains 95% of the posterior probability.



Credible Regions



Bayesian Hypothesis Testing

- Posteriors are proper probability distributions (unlike likelihood curves). So hypotheses can be tested by simply inspecting areas under the curve (e.g. is parameter $x > 1.0$?)
- The Bayesian equivalent of the likelihood ratio statistic is the “**Bayes Factor**”.



That is the worst model I
have ever seen.

Bayesian Model Selection

- The Bayes factor (BF) is the ratio of the probability of model 1 to the probability of model 2, on data D .

$$BF = p(D|M_1) / p(D|M_2)$$

- The models don't have to be nested.
- M_1 and M_2 can represent different models or different regions of parameter space (e.g. $M_1=x<1$ vs $M_2=x>1$).
- Calculating $p(D|M_1)$ is computationally difficult. It is approximated using methods such as “*path sampling*”, “*stepping stone sampling*” or “*HME*”.

Bayesian Model Selection

Bayes factor	Interpretation
<1	M1 is actually worse than M2
1 to 3	Barely worth mentioning
3 to 10	Substantial support for M1
10 to 30	Strong support for M1
30 to 100	Very strong support for M1
>100	Decisive evidence in favour of M1