Bootstrap methods in phylogeny

Strategy

- Assuming only statistical variation
- Carry out measurement “many” times

$$< x > = \frac{1}{N} \sum_{i}^{N} x_i$$

$$\sigma^2 = \frac{1}{N(N-1)} \sum_{i}^{N} (x_i - < x >)^2$$

- Error decreases as number of measurements increase

In fact, there's a huge amount of statistical machinery going on with this.....

Assume the Central Limit Theorem

"If random samples of n observations $y_1, y_2, \ldots y_n$ are drawn from a population of finite mean m and variance s^2, then when n is sufficiently large, the sampling distribution of the sample mean can be approximated by a normal density with mean $m_y = m$ and standard deviation $s_y = s/n^{1/2}$"

THE MOST IMPORTANT THEOREM OF STATISTICS

Consequences of CLT

- Averages taken from *any* distribution (your experimental data) will have a normal distribution
- The error for such an observable will decrease slowly as the number of observations increase

But nobody tells you how big the sample has to be..

- Very often, we are looking at quite complicated *objects,* not just single variables. Even if we assume CLT, then it is not clear how to propagate the uncertainty through to the final objects we are looking at.

- It is not clear when we have a large enough sample, we should do a histogram, but this may not be possible.

What the statistician sees....

- The *probability distribution* rather than the data
- But we just have the data !

The bootstrap method attempts to determine the probability distribution from the data itself, without recourse to CLT. It assesses the accuracy of almost any statistical estimate. It is particularly useful in complicated nonparametric estimation problems, where analytic methods are impractical.

The bootstrap method is not a way of reducing the error ! It only tries to estimate it.

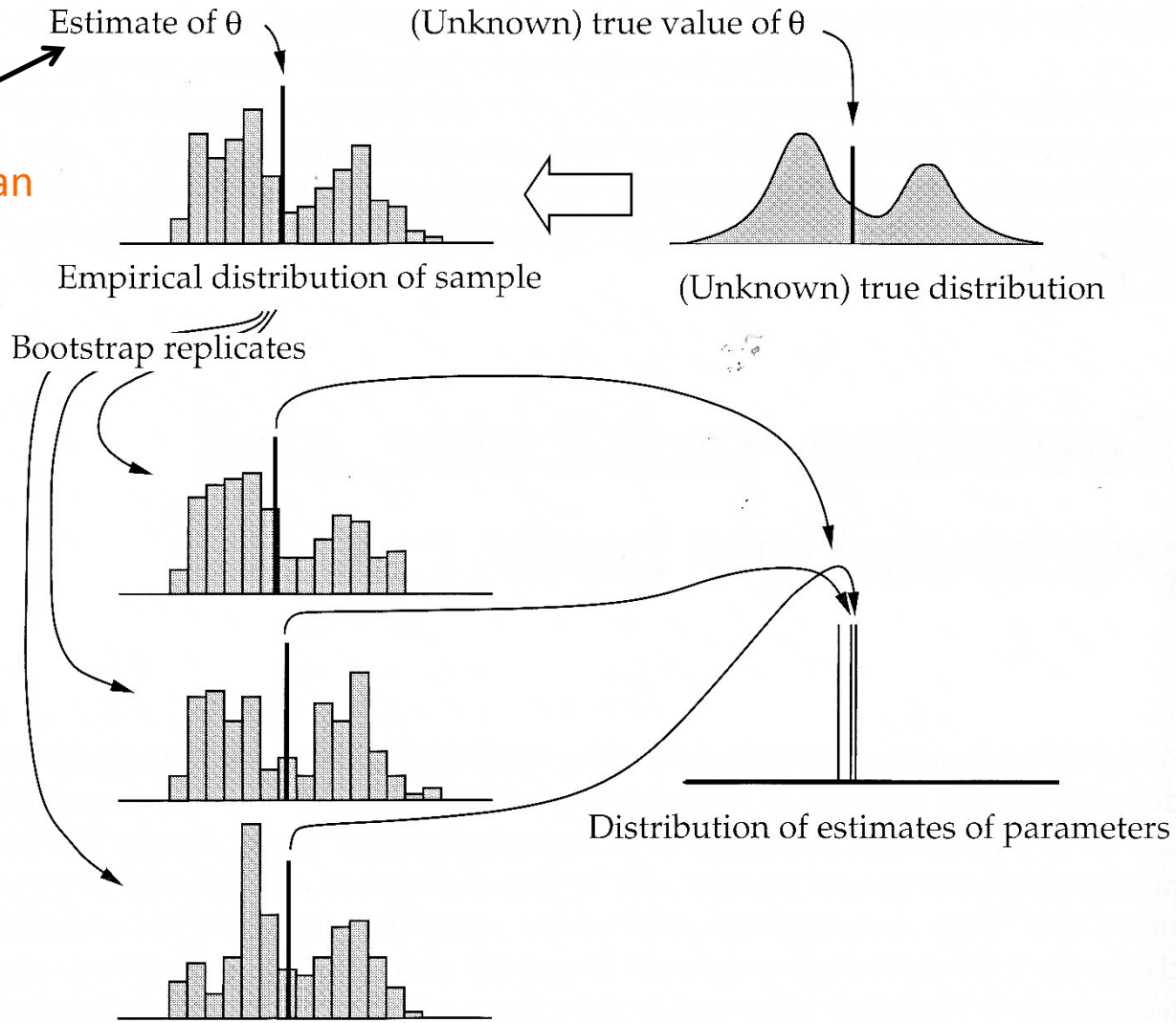
Basic idea of Bootstrap

- Originally, from some list of data, one computes an *object.*
- Create an artificial list by randomly drawing elements from the list of data. *Some elements will be picked more than once.*
- Compute a new object.
- Repeat 100-1000 times and look at the distribution of these objects.

Bootstrap

The distribution of independent data items is taken as an estimation of the unknown true distribution

The population mean is taken as a parameter estimate



Addendum : The Jack-knife

- Jack-knife is a special kind of bootstrap.
- Each bootstrap subsample has all but one of the original elements of the list.
- For example, if original list has 10 elements, then there are 10 jack-knife subsamples.

How many bootstraps ?

- No clear answer to this. Lots of theorems on asymptotic convergence, but no real estimates !
- Rule of thumb : try it 100 times, then 1000 times, and see if your answers have changed by much.
- Anyway have N^N possible subsamples

Is it reliable ?

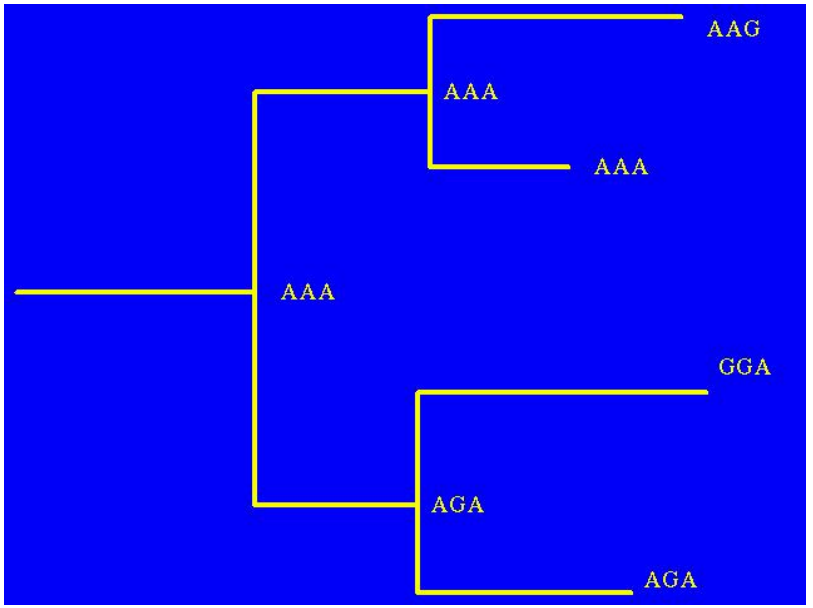
- A very very good question !
- Jury still out on how far it can be applied, but for now nobody is going to shoot you down for using it.
- Good agreement for Normal (Gaussian) distributions, skewed distributions tend to more problematic, particularly for the tails, (bootstrap underestimates the errors).

Case Study : Phylogenetic Trees

Get a multiple sequence alignment

	C1	C2	C3
S1	A	A	G
S2	A	A	A
S3	G	G	A
S4	A	G	A

Construct a Tree using your favourite method (Parsimony, ML, etc..)



How confident are we of this tree ?

- For example, how confident are we that two sequences are in the same clade ?
- I.E. what is the probability distribution of our confidence of the branches ?
- Bootstrap can provide a way of determining this (first thought of by Felsenstein, 1985)

- If we assume there is no correlation between adjoining sites then for a given number of sequences, the tree $\mathcal{T}$ is a function of the set of columns of residues $\mathcal{C} = \{C_1, C_2, \ldots\}$ (i.e. it doesn't matter about their order). In Math Speak :

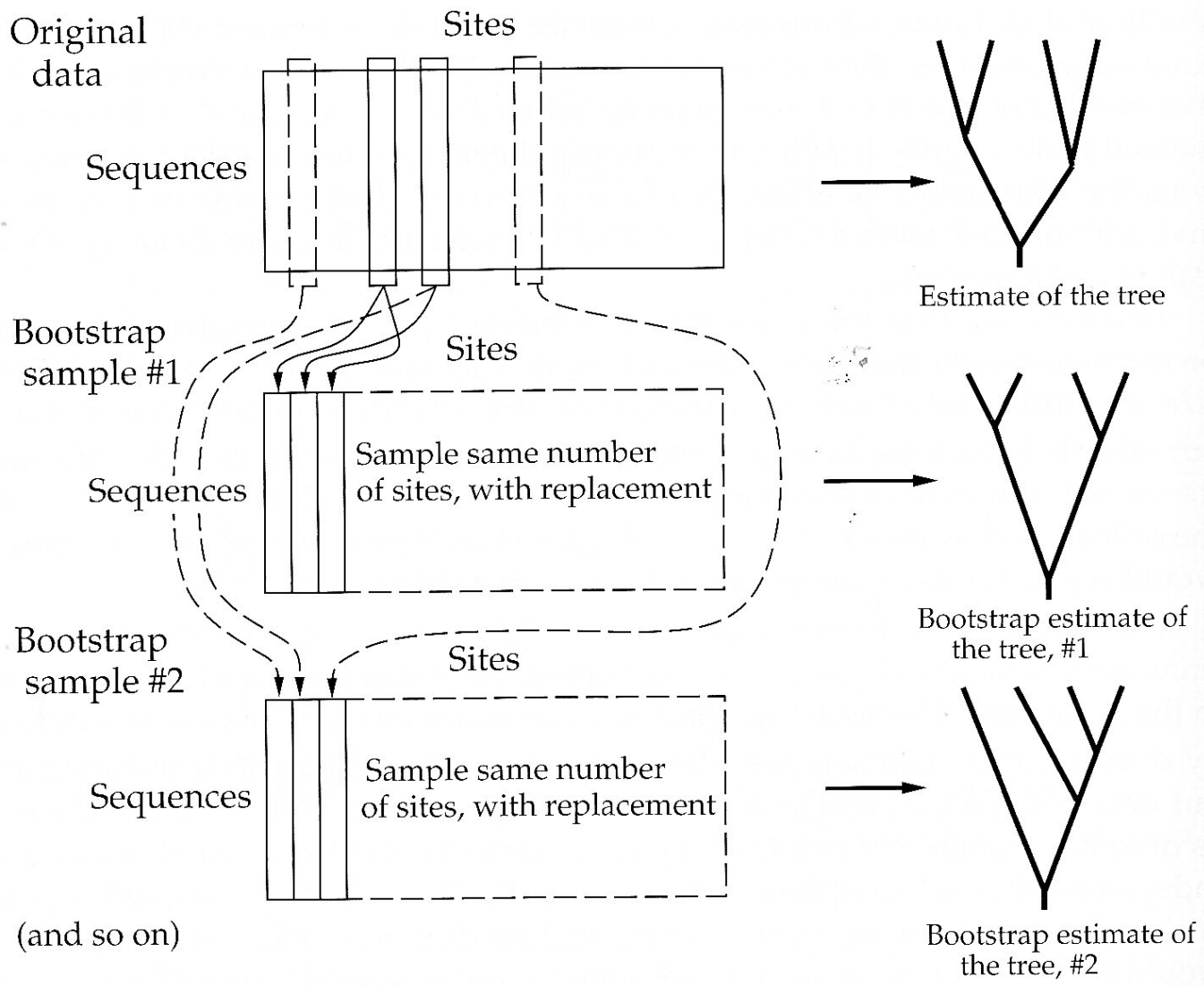
$$\begin{aligned} \mathcal{T} &= \mathcal{T}(C_1, C_2, \ldots C_i, \ldots C_j, \ldots) \\ &= \mathcal{T}(C_1, C_2, \ldots C_j, \ldots C_i, \ldots) = \mathcal{T}(\mathcal{C}) \end{aligned}$$

- So, construct an ensemble of bootstrap subsamples

$$\begin{aligned} \mathcal{C}_1 &= B_o[1; \mathcal{C}] \\ \mathcal{C}_2 &= B_o[2; \mathcal{C}] \\ \ldots &\quad \ldots \end{aligned}$$

and hence compute the equivalent trees $\mathcal{T}_1, \mathcal{T}_2, \ldots$.

Bootstrap for phylogeny



Having created an ensemble of phylogenetic trees, one can elucidate the statistical frequency of various features of the tree.

E.G. Do two sequences lie in the same clade ?

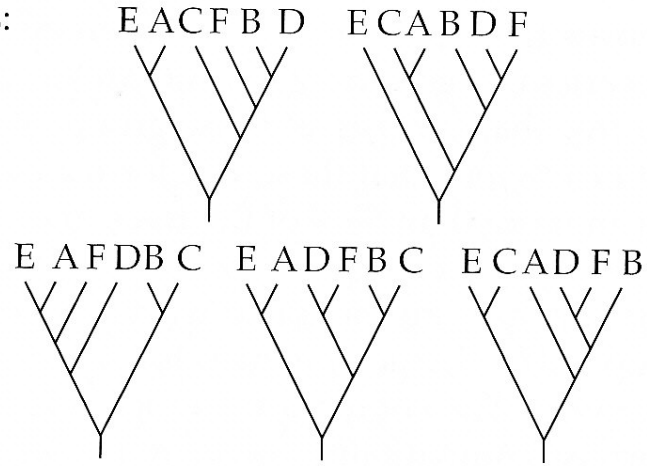
$$P(Clade|S_i, S_j) = \frac{1}{N_{boot}} \sum_{I}^{N_{boot}} Clade[T(S_i, S_j)]$$

Can this be used for statistical significance ?

This is very much an open question !!!!

(Be cautious, ……)

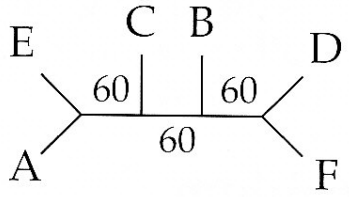
Trees:



Number of times each partition of species is found:

AE | BCDF	3
ACE | BDF	3
ACEF | BD	1
AC | BDEF	1
AEF | BCD	1
ADEF | BC	2
ABDF | EC	1
ABCE | DF	3

Majority–rule consensus tree of the unrooted trees:



The 60% corresponds to 3 times over a total of 5

A set of 5 trees and their consensus tree, with the percentage of support for each interior branch shown. Note that the topology of the consensus tree is not identical to the topology of any of the 5 trees. The 5 trees are considered unrooted in the computation.

Quelques references bibliographiques