Bootstrap confidence levels for phylogenetic trees

B. Efron, E. Halloran, and S. Holmes, 1996

Following “Confidence limits on phylogenies: an approach using the bootstrap”, **J. Felsenstein**, 1985

I. Short review of the bootstrap method
[Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap*;
Saharon's course "Bootstrap and Resampling Methods"]

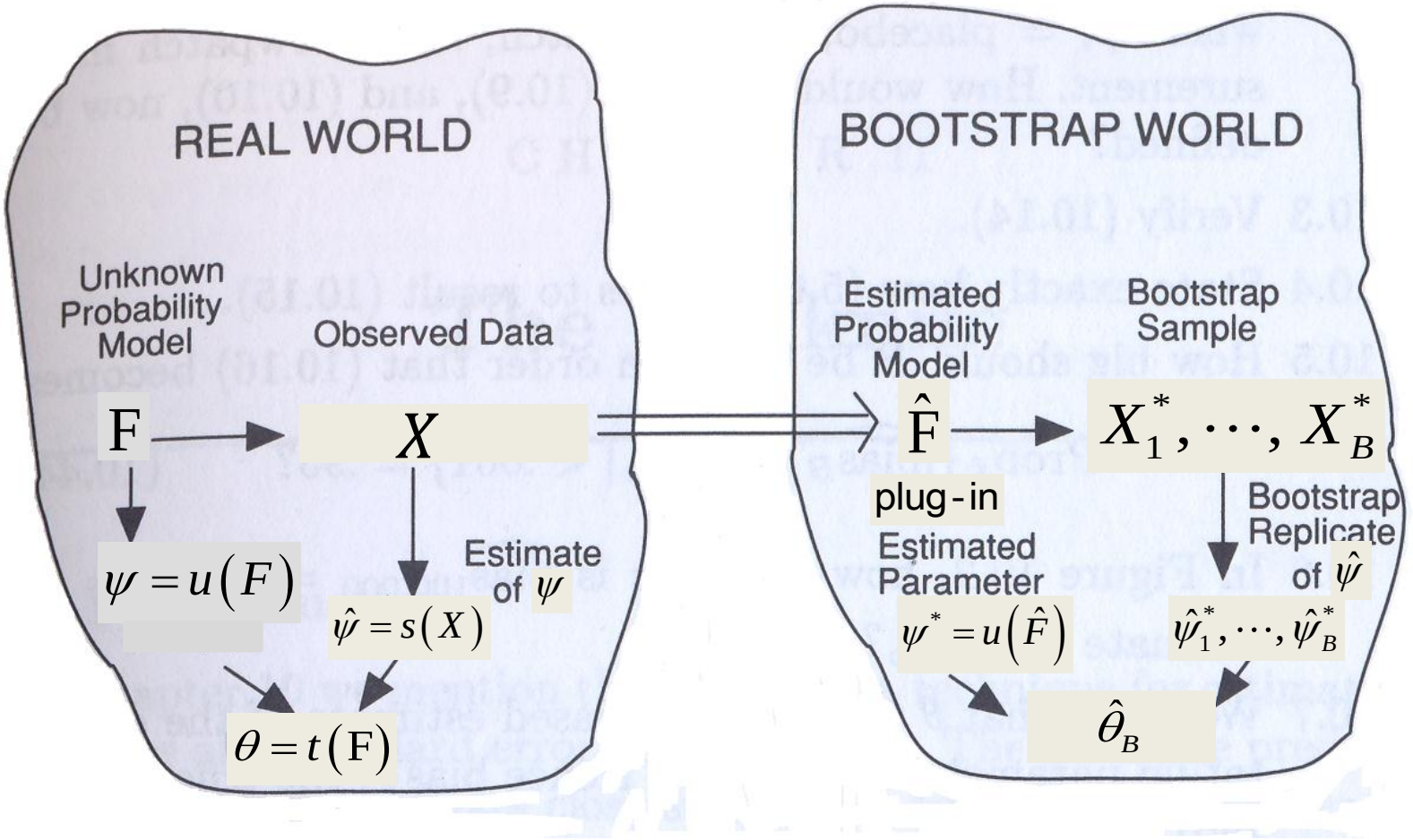
II. Short review of phylogenetic trees
[Saharon's course "Topics in Statistical Genetics"]

III. Felsenstein's method for computing confidence levels on monophyly of groups in phylogenetic trees using the bootstrap, and critiques of it
[Felsenstein, J. (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap;
Saharon's course "Bootstrap and Resampling Methods"]

IV. Efron's justification of Felsenstein's method and corrections of it
[**Efron, B., Halloran, E. & Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees**;
Efron, B. (1982) The jackknife, the bootstrap and other resampling plans;
Efron, B. (1987) Better Bootstrap Confidence Intervals]

I. Short review of the bootstrap method

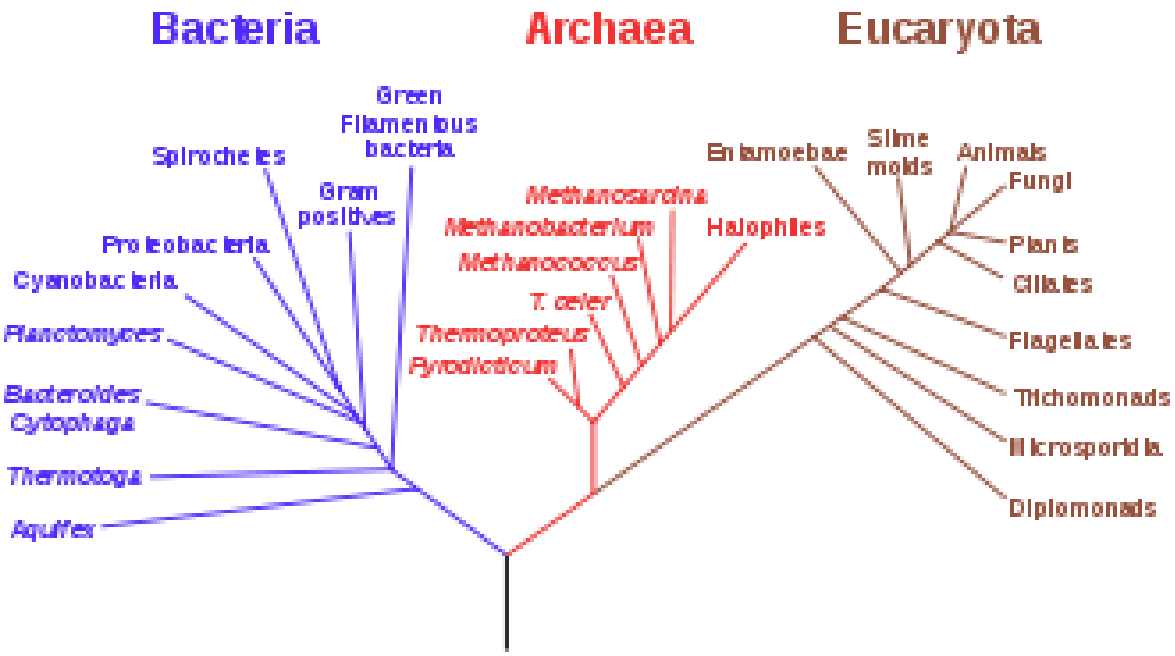
- Population distribution function F, some parameter or feature $\psi = u(\mathrm{F})$
- A random sample $X = (x_1, \cdots, x_n)$ from this population
- Estimator $\hat{\psi} = \mathrm{s}(X)$
- What is the estimator's quality? We want to estimate some parameter that characterizes it, $\theta = t(\mathrm{F})$, which may be a function of the unknown parameter ψ.


REAL WORLD
Unknown Probability Model
Observed Data
F
X
$\psi = u(F)$
Estimate of ψ
$\hat{\psi} = s(X)$
$\theta = t(\mathrm{F})$
BOOTSTRAP WORLD
Estimated Probability Model
Bootstrap Sample
$\hat{\mathrm{F}}$
$X_1^*, \cdots, X_B^*$
plug-in
Estimated Parameter
Bootstrap Replicate of $\hat{\psi}$
$\psi^* = u(\hat{F})$
$\hat{\psi}_1^*, \cdots, \hat{\psi}_B^*$
$\hat{\theta}_B$

II. Short review of phylogenetic trees

- Every organism carries sequences of DNA, and each species is characterized by its typical sequences.
- The history of all species begins with one ancestor, whose offspring developed various mutations. Throughout history species were formed through accumulation of mutations.
- We want to build a tree to map the history and relationship between species. All species existing today are leaves in this tree.
- The real tree is rooted, since there was only one species in the beginning.

Phylogenetic Tree of Life

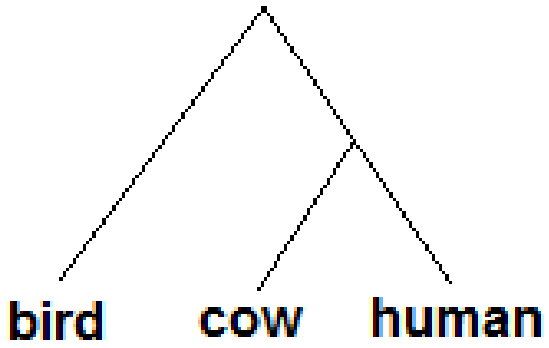


Bacteria - a large domain of prokariotic microorganisms (**Prokariots** - organisms whose cells lack a cell nucleus (karyon)).

Archaea - another domain of prokariots, all single-celled.

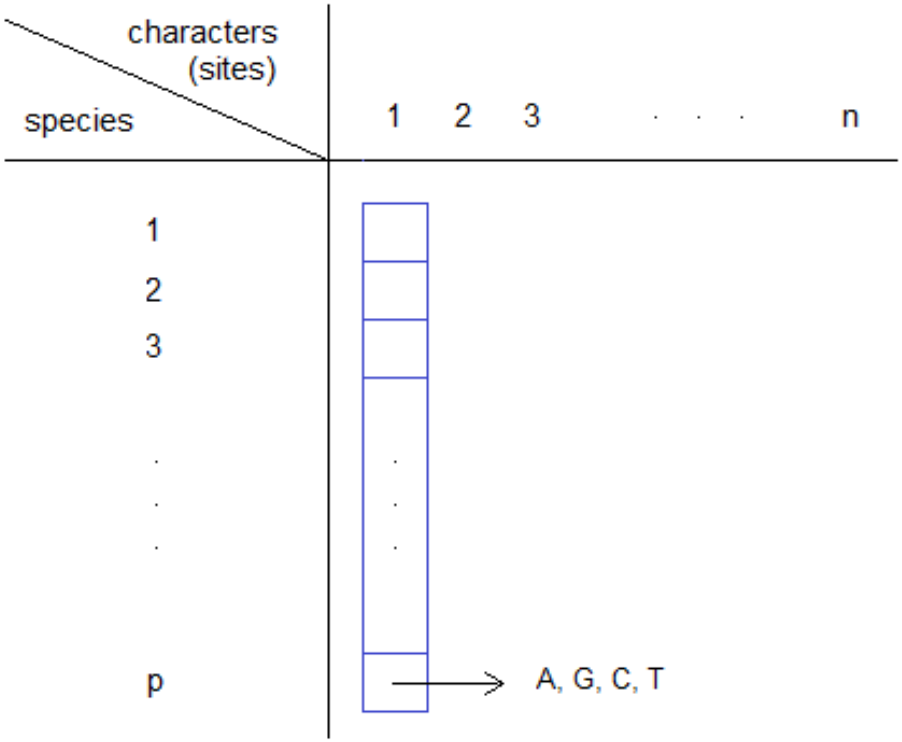
Eucaryota - organisms whose cells do have a nucleus.

- If we take sequences from very different species, with high probability all nucleotides can be described by the same tree.



It means that for every site, the convergence of (human, cow) is closer than of (human, bird) and of (cow, bird). This is probable given the history.

- Table of sequences:



- The simplest metrics between species: **Hamming distance** - the number of sites at which the two species differ.

- Tree building methods:

  1. Neighbour joining:
  Connect the two closest species, and so on.

  2. Maximum parsimony:
  Minimize the total number of mutations in the tree.
  Good for many thousands of species.

3. Maximum likelihood:

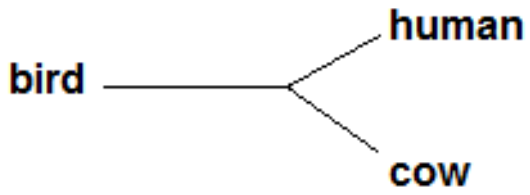
Parameters:

- Tree topology
- Mutation model
- Edge lengths

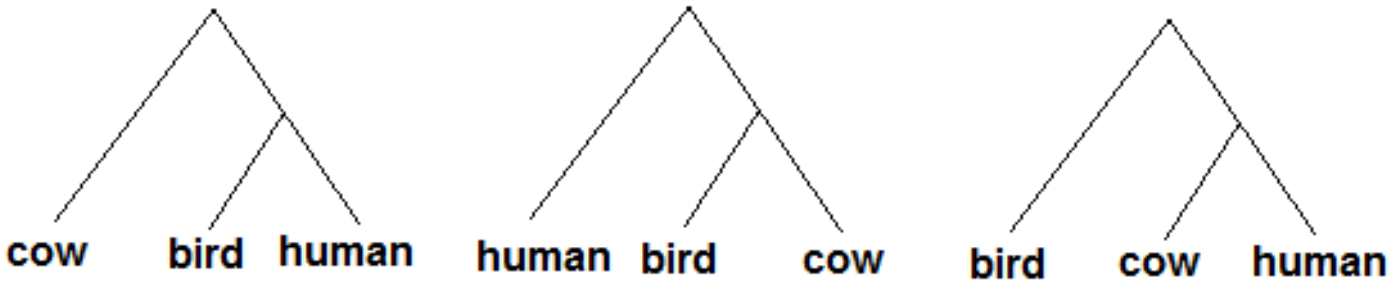
Usually assume that the sites are independent.

Good for a few tens of species.

- None of these models gives information about the root's location.


human
bird
cow

?


cow
bird
human
human
bird
cow
bird
cow
human

- Methods for placing the root:

  1. Molecular clock:

  Mutations occur in a constant rate over history.

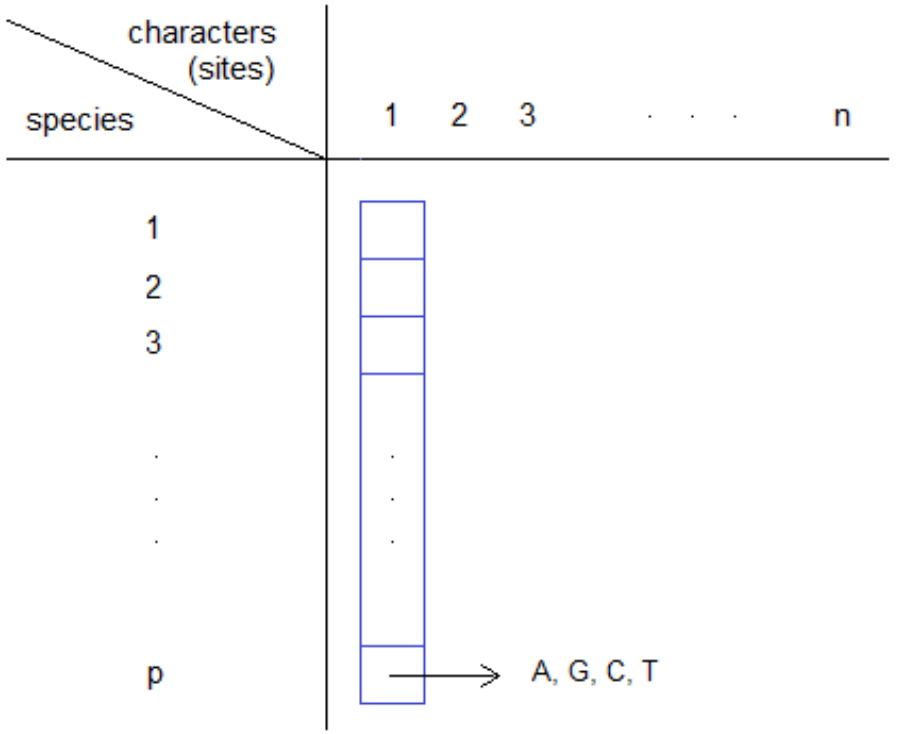
  Choose the most distant point from the leaves, according to some metrics (e.g. the point with biggest minimal distance from a leaf).

  2. Outgroup: Add a species that is surely separated from all, and “hang” the tree from the edge leading to it.

III. Felsenstein's method

- a **monophyletic** group is a group of species which consists of all the species that are descended from some edge in the rooted tree, and only of them.
- A rooted tree is a series of arguments about the monophyly of groups in the tree.
- Use bootstrap to build phylogenetic trees and test hypotheses about the monophyly of groups
- The researcher's main interest is to decide whether or not a group of species is monophyletic.

- Data table:



- Bootstrap across the characters: sample n columns with replacement.

- We get a bootstrap sample that is a multinomial r.v. : each of the possible columns has a probability that is its frequency in the original sample of columns.
- Choose some optimality criterion for tree building (e.g. maximum parsimony).
- Build a tree based on each of the bootstrap samples.
- The estimated tree(s) is built according to the “majority rule”: the groups that are monophyletic in it are the ones that where monophyletic in most of the bootstrap trees.

- S - some group of species.
- H_0: S is *not* monophyletic
- Reject H_0 If S is monophyletic in at least 95% of the bootstrap trees, i.e. if it is *not* monophyletic in less than 5% of the bootstrap trees.
- This is equivalent to building a one-sided percentile confidence interval for the indicator

$$\psi = \begin{cases} 1, & S \text{ is monophyletic} \\ 0, & \text{otherwise} \end{cases}$$

- For each bootstrap sample ($b = 1, \cdots, B$):

$$\hat{\psi}_b^* = \begin{cases} 1, & S \text{ is monophyletic according to sample } b \\ 0, & \text{otherwise} \end{cases}$$

And $\mathrm{CI}_{\text{pct, 0.95}} = \left[\hat{\psi}^*_{(0.05)}, 1\right]$.

- If the 5th percentile of $\left\{\hat{\psi}_b^*\right\}_{b=1}^{B}$ is 1 (i.e., the lower bound of the bootstrap percentile confidence interval is 1) , Felsenstein rejects H_0 and says that the group is monophyletic.
- The proportion of bootstrap trees where S is monophyletic is Felsenstein's **confidence** in the monophyly of the group. Efron denotes it by $\tilde{\alpha}$ $\left(= 1 - \left[\text{Felsenstein's p-value}\right]\right)$.

- Data set used by Efron to demonstrate the use of Felsenstein's method: 11 malaria species, 221 sites.

  The first 20 columns of the data matrix X :

	Site:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Species	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	Pre (Chimp)	C	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
2	Pme (Lizard)	T	C	T	A	A	A	A	G	A	T	T	A	T	A	T	A	G	A	T	A
3	Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T	T	A	A	A	T	T
4	Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
5	Pbe (Rodent)	T	T	T	A	A	G	A	A	A	A	T	T	T	A	T	A	A	A	T	A
6	Plo (Bird)	T	T	T	A	A	G	A	A	A	A	C	T	C	A	C	A	A	A	T	C
7	Pfr (Monkey)	C	T	T	A	A	G	A	A	G	A	T	T	C	T	T	A	G	G	A	A
8	Pkn (Monkey)	C	T	T	A	A	G	A	A	A	G	T	T	C	T	T	A	G	A	T	A
9	Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
10	Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	T	C	T	C	G	G	A	T	A
11	Pga (Bird)	T	T	T	A	A	G	A	A	A	A	T	T	T	T	C	A	A	A	T	C

- Tree building algorithm: distance matrix $\hat{D}$ between species (221 x 221)

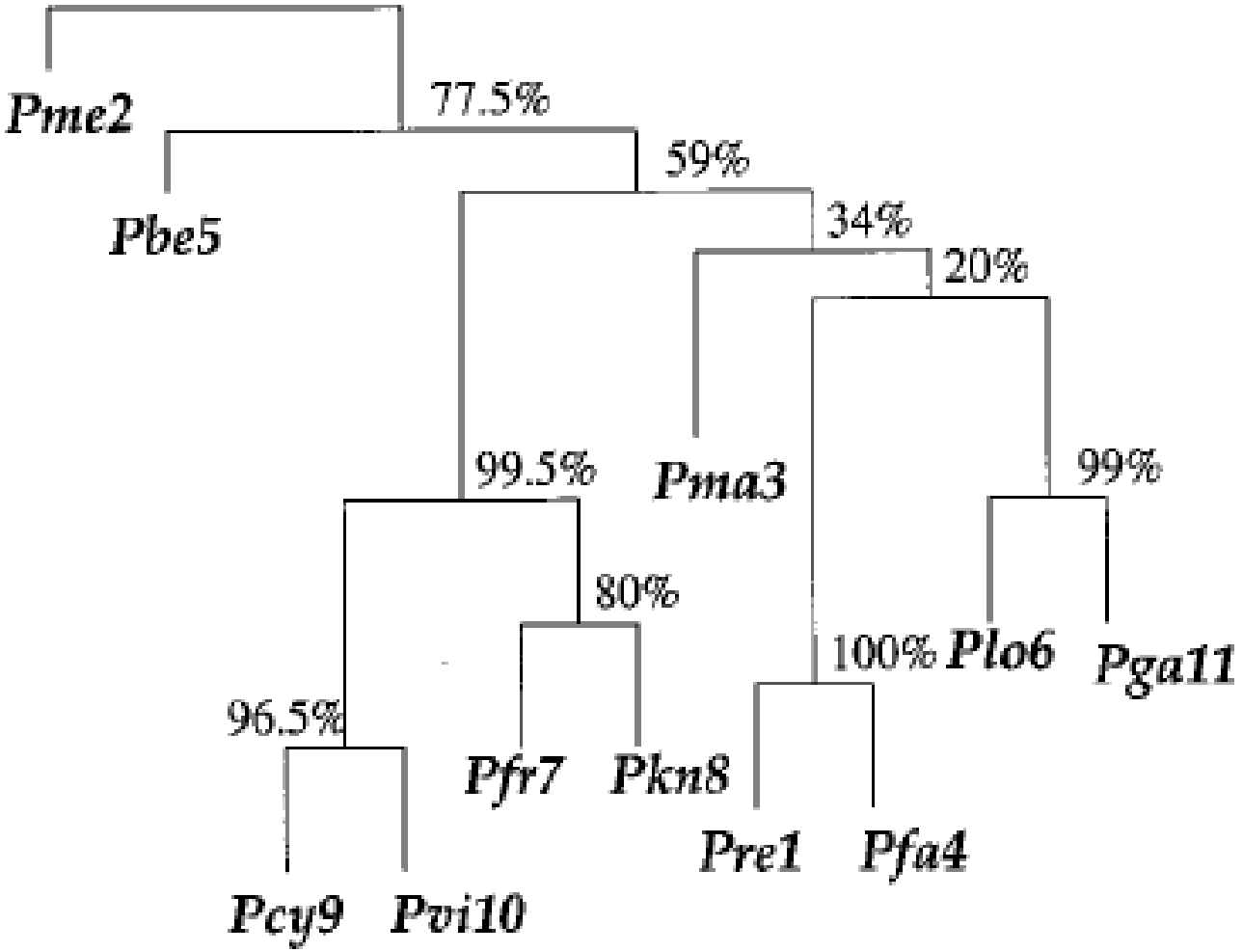
$$X \rightarrow \hat{D} \rightarrow \hat{\text{tree}}$$

- Proceeding with Felsenstein's method:

  $B = 200$ bootstrap samples were generated, and bootstrap trees were built:

$$X^* \rightarrow \hat{D}^* \rightarrow \hat{\text{tree}}^*$$

- For every monophyletic group in the original tree, the proportion of bootstrap trees where this group is monophyletic (Felsenstein's confidence) was calculated.
- For example: the 9-10 clade appeared (as monophyletic) in 193 of the 200 bootstrap trees, so its confidence value is 0.965.


Pme2
77.5%
Pbe5
59%
34%
20%
99.5%
Pma3
99%
80%
100%
Plo6
Pga11
96.5%
Pfr7
Pkn8
Pre1
Pfa4
Pcy9
Pvi10

What are the main problems with Felsenstein's method:

1) With respect to this course?
2) As put by Efron?
3) According to other critiques, such as Hillis and Bull?

1) With respect to this course:

- This is not correct hypothesis testing:

$$\psi = \begin{cases} 1, & S \text{ is monophyletic} \\ 0, & \text{otherwise} \end{cases}$$

H_0: S is *not* monophyletic, i.e. $\psi = 0$.

H_1: S *is* monophyletic, i.e. $\psi = 1$.

Test statistic: $\hat{\psi} = s(X)$

We need to create a *null* bootstrap world, where S is *not* monophyletic, i.e. $\psi = 0$, and use it to estimate $\mathrm{P}_{H_0}\left(s(X) \geq s(X)_{\text{obs}}\right)$ by calculating

$$\hat{\mathrm{P}}_{\mathrm{F}_0}\left(s(X) \geq s(X)_{\mathrm{obs}}\right) = \hat{\mathrm{ASL}}_{\mathrm{boot},\,B} = \frac{\#\left\{s\left(X_b^*\right) \geq s(X)_{obs}\right\}}{B}$$

This is the proportion of bootstrap samples that give a tree in which S is monophyletic, when F_0, the distribution function of bootstrap samples X^*, is in H_0 and is as similar as possible to F (the distribution of X in the real world).

Instead, Felsenstein resamples from a specific distribution in H_1 – the empirical distribution of X, that gave us a tree in which S *is* monophyletic. This Is not valid hypothesis testing.

- Bootstrap percentile confidence interval cannot be justified, for example because the monophyly indicators ψ and $\hat{\psi}$ can only get 0 or 1, so there cannot exist a monotone g s.t.

$$\mathrm{g}(\hat{\psi}) - g(\psi) \sim \mathrm{N}(0, \gamma^2).$$

2) As put by Efron:

- The use of the (nonparametric) bootstrap to estimate the parameter itself (ψ, the indicator for the monophyly of S, by estimating the tree), instead of the quality of the estimator (bias, variance etc.).
- We *want* to infer our confidence that S is *truly* monophyletic given that we estimated it as monophyletic.
- But we are *actually estimating* the probability of *inferring* monophyly, assuming that our original sample is a true description of the real tree (in the sense that S is indeed monophyletic).

In other words:

- We *want* to decide how confident (confidence = 1 - p-value) we are that $\psi = 1$ given that our estimate is $\hat{\psi} = 1$. This is equivalent to building a valid percentile plug-in CI: $\text{CI}_{\text{pct, plug-in, 0.95}} = \left[\hat{\psi} - \left(\hat{\psi}^*_{(0.95)} - \hat{\psi}\right), 1\right] = \left[2\hat{\psi} - \hat{\psi}^*_{(0.95)}, 1\right]$
- But we are *actually estimating* the probability to get $\hat{\psi} = 1$ when $\psi = 1$. This is equivalent to building the less valid percentile CI:

$$\text{CI}_{\text{pct, 0.95}} = \left[\hat{\psi}^*_{(0.05)}, 1\right]$$

3) According to other critiques, such as Hillis and Bull:

Felsenstein's confidence values are consistently too conservative (i.e., biased downward) as an assessment of the tree accuracy.

IV. Efron's justification and corrections

Efron, Halloran & Holmes:

a) Felsenstein's method provides a reasonable first approximation to the actual confidence levels of the observed clades. This is an answer to problem (2).

b) Felsenstein's confidence assessment is *not* consistently biased downward. This is an answer to problem (3).

c) Suggest a more complex method to give better assessments of confidence.

a) Efron's answer to problem (2): Why is Felsenstein's confidence justified?

- The rationale: The bootstrap samples come from a multinomial model:
- In the malaria example, there are 11 species, so there are $K = 4^{11} - 4$ possible column vectors for x. Denote them $y_1, \cdots, y_K$.
- Suppose each observed column of X is independently chosen from these vectors, with some probability π_k to choose y_k ($\sum_{k=1}^{K} \pi_k = 1$). These probabilities depend on the population of sites (the true tree).
- Denote $\pi = (\pi_1, \cdots, \pi_K)$.

- X can be characterized by the proportion of its $n = 221$ columns equaling each possible y_k :

$$\hat{\pi}_k = \frac{\#\{\text{columns of } X \text{ equaling } y_k\}}{n}$$

$$\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_K)$$

- $\hat{D}$ is a function of the original observed proportions $\hat{\pi}$, so the tree-building algorithm can be described as

$$\hat{\pi} \rightarrow \hat{D} \rightarrow \text{trêe}$$

- We can add the group of interest's monophyly indicator to the process:

$$\hat{\pi} \rightarrow \widehat{D} \rightarrow \hat{\text{tree}} \rightarrow \hat{\psi}$$

- In a similar way (although possibly in an opposite causal order – the tree comes first in reality), the vector of true probabilities π gives the true distance matrix and the true tree (that we assume we would get by applying our algorithm to the true distance matrix):

$$\pi \rightarrow D \rightarrow \text{tree} \rightarrow \psi$$

(where $D_{ij} = \sqrt{\sum_{k=1}^{K} \pi_k \left(y_{ki} - y_{kj} \right)^2}$) .

- The proportions of columns in a bootstrap sample are:

$$\hat{\pi}_k^* = \frac{\#\left\{\text{columns of } X^* \text{ equaling } y_k\right\}}{n}$$

$$\hat{\pi}^* = \left(\hat{\pi}_1^*, \cdots, \hat{\pi}_K^*\right)$$

- And we get a bootstrap tree:

$$\hat{\pi}^* \rightarrow \hat{D}^* \rightarrow \hat{\text{tree}}^*$$

(and $\hat{\psi}^*$).

A schematic picture of the space of possible π's:

- We hope to have trêe = tree .

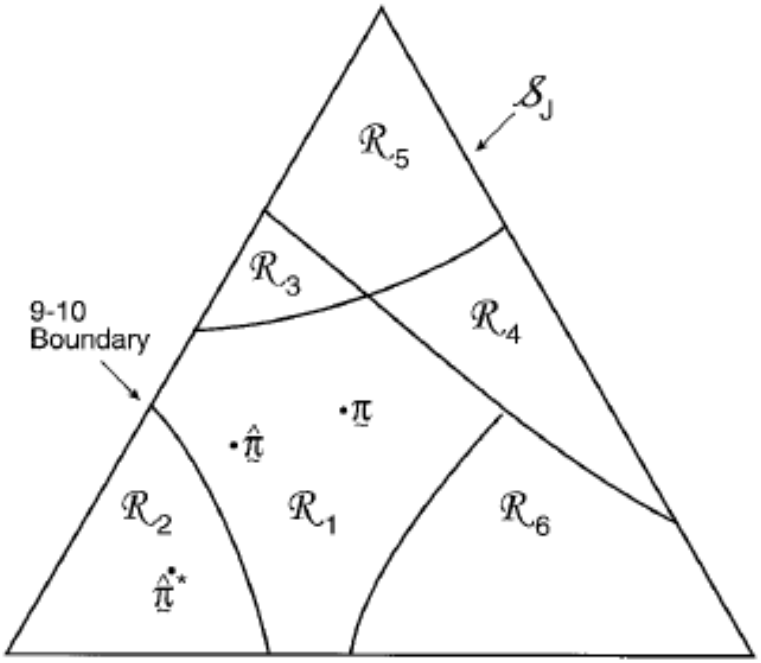


FIG. 3. Schematic diagram of tree estimation; triangle represents the space of all possible π vectors in the multinomial probability model; regions $\mathcal{R}_1$, $\mathcal{R}_2$. . . correspond to the different possible trees. In the case shown π and $\hat{\pi}$ lie in the same region so TREE = $\widehat{\text{TREE}}$, but $\hat{\pi}^*$ lies in a region where $\widehat{\text{TREE}}^*$ does not have the 9-10 clade.

- It is not clear why we can use the distribution of $\text{trêe}^* \,|\, \text{trêe}$ to infer our confidence in $\text{tree} \,|\, \text{trêe}$.
- Or in terms of the monophyly indicator, why we can decide how confident we are that $\psi = 1$ given that our estimate is $\hat{\psi} = 1$ by estimating the probability that $\hat{\psi} = 1$ given that $\psi = 1$.

The answer is in using a Bayesian approach:

- The bootstrap distribution of $\hat{\text{tree}}^* \,|\, \hat{\text{tree}}$ is almost the same as the *posterior* distribution of $\text{tree} \,|\, \hat{\text{tree}}$ if we begin with an "uninformative" (i.e., uniform) *prior* density π.

An explanation can be found in Efron's booklet "*The Jackknife, the Bootstrap and other resampling plans*" (1982).

- Assume a discrete sample space. In our phylogenetic case, this is the space of all possible columns ($K = 4^{11} - 4$).
- π_k - the probability that a column in a sample will be identical to the possible column y_k .
- The observed frequency: $\hat{\pi}_k = \frac{\#\{x_i \text{ equaling } \mathrm{y_k}\}}{n}$

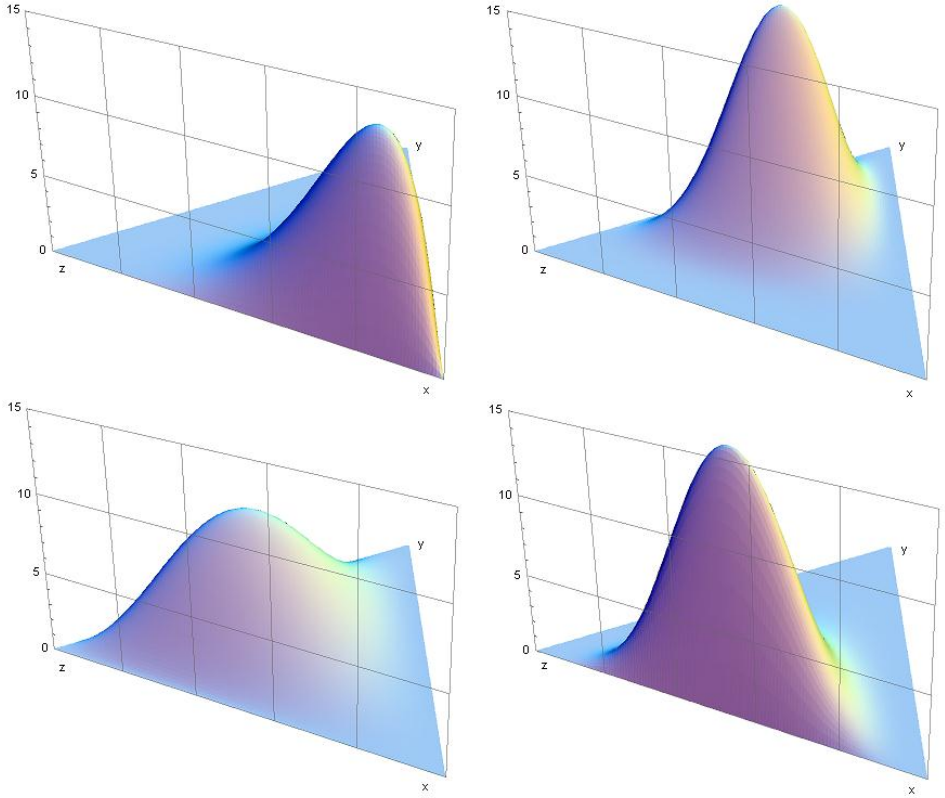
- The observed frequencies: $\hat{\pi} = (\hat{\pi}_1, \cdots, \hat{\pi}_K)$
- The real probabilities: $\pi = (\pi_1, \cdots, \pi_K)$

- Choose as the prior distribution of π (the real probabilities – the parameter we want to estimate) a $K-$dimentional Dirichlet distribution with parameter α, which has to be a vector of K positive numbers:

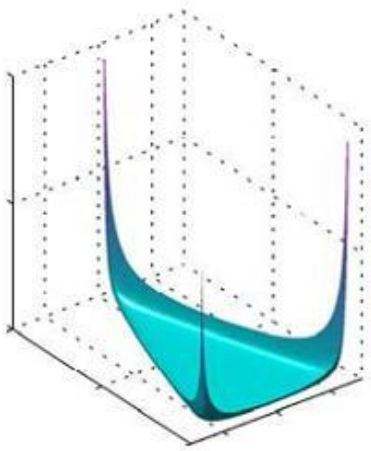
  $\alpha = (\alpha_1, \cdots, \alpha_K)$.

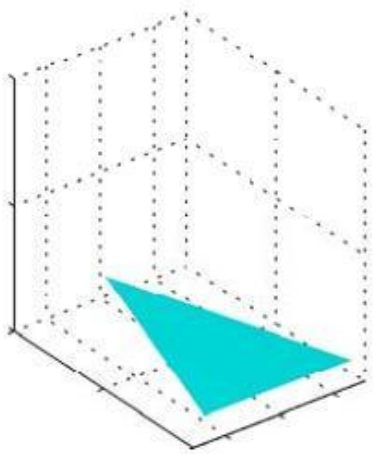
- We choose $\alpha = (a, \cdots, a) = a \cdot \mathbf{1}$ to get a symmetric distribution.

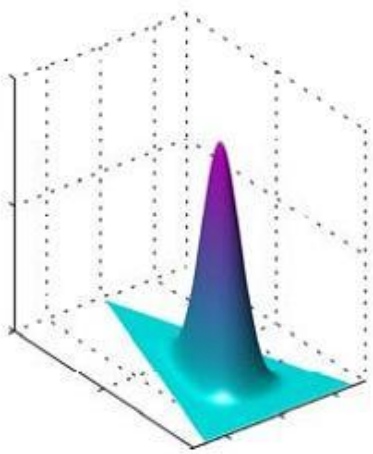
- From Wikipedia: Several images of probability densities of the Dirichlet distribution with K=3. Clockwise from top left: α = (6,2,2), (3,7,5), (6,2,6), (2,3,4).



- Left plot: $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$
- Center plot: $\alpha_1 = \alpha_2 = \alpha_3 = 1$
- Right plot: $\alpha_1 = \alpha_2 = \alpha_3 = 10$







- So apriori $\pi = (\pi_1, \cdots, \pi_K) \sim \mathrm{Dir}_K(\alpha)$ with the density function

$$f(\pi_1, \cdots, \pi_K) = \begin{cases} \dfrac{1}{\mathrm{B}(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, & \pi_1 + \cdots + \pi_K = 1,\ \pi_1, \cdots, \pi_K > 0 \\ 0, & \text{otherwise} \end{cases} \quad \begin{matrix} \alpha = a \cdot \mathbf{1} \\ \propto \end{matrix} \quad \prod_{k=1}^{K} \pi_k^{a-1}$$

- From the sample we have the observed frequencies $\hat{\pi}$.
- The likelihood function $\hat{\pi}\left|\pi\right.$ is multinomial:

$$n\hat{\pi}\left|\pi\right. \sim \text{Mult}\left(n,\pi\right)$$

$$\hat{\pi}\left|\pi\right. \sim \frac{\text{Mult}\left(n,\pi\right)}{n}$$

$$\text{P}\left\{\hat{\pi}\left|\pi\right.\right\} = \begin{cases} \dfrac{\prod_{k=1}^{K}\left(n\hat{\pi}_k\right)!}{n!}\prod_{k=1}^{K}\pi_k^{n\hat{\pi}_k}, & \sum_{k=1}^{K} n\hat{\pi}_k = n, \quad n\hat{\pi}_k \in \text{Z} \\ 0, & \text{otherwise} \end{cases}$$

- The Dirichlet distribution is self conjugate with regard to multinomial likelihood, and the posterior distribution of π is

$$\pi \left| \hat{\pi} \sim \mathrm{Dir}_K \left(a \cdot \mathbf{1} + \mathrm{n}\hat{\pi} \right) \right.$$

  which means that we add to each parameter the number of times that the corresponding column appears in the sample X.

- If we minimize the effect of the prior "knowledge" by taking $a \to 0$, we will get

$$\pi \left| \hat{\pi} \sim \mathrm{Dir}_K \left(\mathrm{n}\hat{\pi} \right) \right.$$

- The distribution of the observed frequencies in a bootstrap sample,

$$\hat{\pi}^* \left| \hat{\pi} \sim \frac{\mathrm{Mult}\left(n, \hat{\pi}\right)}{n} \right.$$

is a good approximation for the limit of the posterior distribution $\pi \left| \hat{\pi} \sim \mathrm{Dir}_K \left(\mathrm{n} \hat{\pi} \right) \right.$.

- Therefore we have a basis to think that Felsenstein's confidence level is a good approximation for the confidence level we try to estimate.

b) Efron's answer to problem (3): Why is Felsenstein's confidence assessment *not* consistently biased downward?

- To better explain the problem in Felsenstein's method which causes the apparent "bias", Efron uses a simpler example: A normal model (instead of multinomial) and parametric bootstrap (instead of nonparametric).

A simpler model: $\boldsymbol{x} \sim \mathrm{N}_2(\boldsymbol{\mu}, \mathrm{I})$

$$\boldsymbol{\mu} = (\mu_1, \mu_2)$$

$$\hat{\boldsymbol{\mu}} = \boldsymbol{x}$$

- The $\boldsymbol{\mu}$ - plane is divided into regions $\Re_1, \Re_2$.
- $\hat{\boldsymbol{\mu}}$ lies in $\Re_1$, and we wish to assign a confidence value to the event that $\boldsymbol{\mu}$ itself lies in $\Re_1$.
- In our terms:

  H_0: $\boldsymbol{\mu} \in \Re_2$

  H_1: $\boldsymbol{\mu} \in \Re_1$

  confidence = 1 - p-value

- Parametric bootstrap:

$$\boldsymbol{x}^* \sim \mathrm{N}_2(\hat{\boldsymbol{\mu}}, \mathrm{I})$$

- Felsenstein's confidence value:

$$\tilde{\alpha} = \mathrm{P}_{\hat{\mu}}\left\{\hat{\boldsymbol{\mu}}^* \in \Re_1\right\}$$

- As in the multinomial case, it can be justified by using the Bayesian argument.

Why is $\tilde{\alpha}$ a reasonable assessment of the confidence that $\boldsymbol{\mu} \in \Re_1$?

- As in the multinomial model, $\tilde{\alpha}$ is the posterior probability that $\boldsymbol{\mu} \in \Re_1$ given that $\hat{\boldsymbol{\mu}} \in \Re_1$ when we assume apriori that $\boldsymbol{\mu}$ can be anywhere in the plane with equal probability.
- The posterior distribution: $\boldsymbol{\mu} \,|\hat{\boldsymbol{\mu}} \sim \mathrm{N}_2(\hat{\boldsymbol{\mu}}, \mathrm{I})$, exactly the distribution of $\hat{\boldsymbol{\mu}}^* \,\big|\hat{\boldsymbol{\mu}}$.

- $\hat{\boldsymbol{\mu}} - \boldsymbol{\mu},\ \hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}} \sim \mathrm{N}_2(\mathbf{0}, \mathrm{I})$, but $\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu} \sim \mathrm{N}_2(\mathbf{0}, 2\mathrm{I})$.

According to Efron, this generates the "bias" for which the method was criticized.

- In the phylogeny case, the fact is that the probability that $\hat{\text{tree}}^* = \text{tree}$ is usually less than the probability that $\hat{\text{tree}} = \text{tree}$.

- But Efron also gives a perhaps more convincing explanation for the alleged bias.

- Two possible examples:

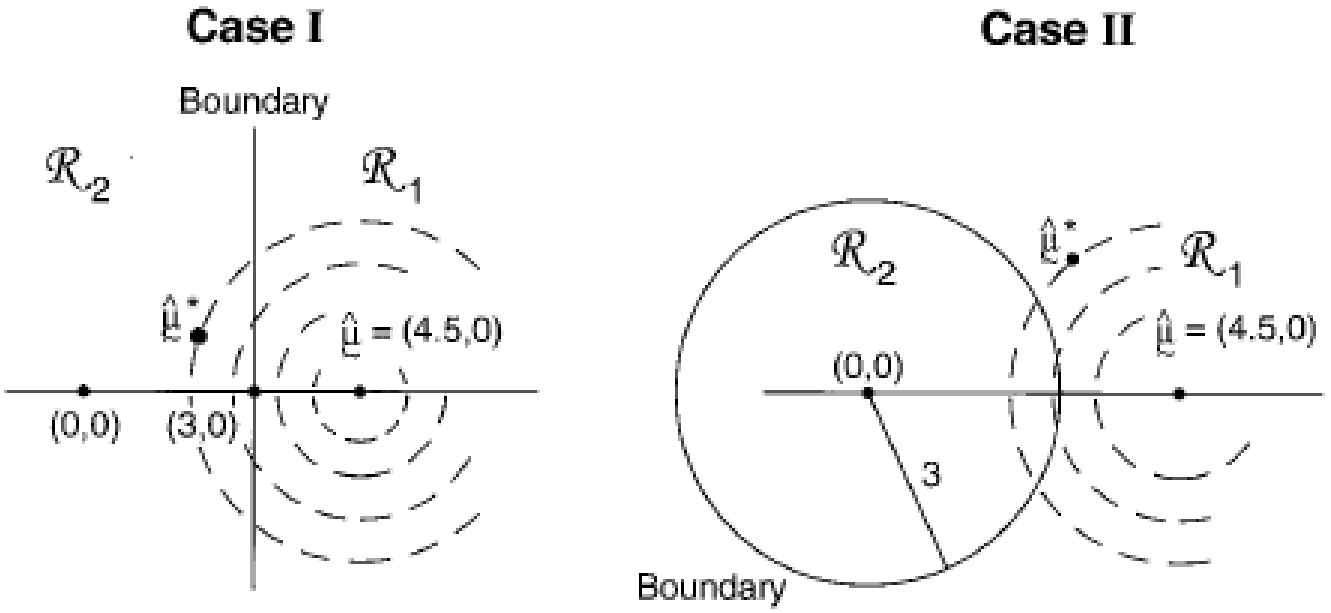


FIG. 4. Two cases of the simple normal model; in both we observe $\hat{\boldsymbol{\mu}} = (4.5, 0) \in \mathcal{R}_1$, and wish to assign a confidence value to $\boldsymbol{\mu} \in \mathcal{R}_1$. Case I, $\mathcal{R}_2$ is the region $\{\mu_1 \leq 3\}$. Case II, $\mathcal{R}_2$ is the region $\{\|\boldsymbol{\mu}\| < 3\}$. The dashed circles indicate bootstrap sampling $\hat{\boldsymbol{\mu}}^* \sim N_2(\hat{\boldsymbol{\mu}}, I)$.

- Felsenstein's confidence value:

$$\tilde{\alpha} = \mathrm{P}_{\hat{\mu}} \left\{ \hat{\boldsymbol{\mu}}^* \in \Re_1 \right\}$$

- Calculated theoretically:

$$\tilde{\alpha}_{\mathrm{I}} = 0.933, \quad \tilde{\alpha}_{\mathrm{II}} = 0.949$$

- How would we assign a confidence level to $\boldsymbol{\mu} \in \Re_1$?
- In our terms:

  H_0: $\boldsymbol{\mu} \in \Re_2$

  H_1: $\boldsymbol{\mu} \in \Re_1$

  confidence = 1 - p-value

- A more customary way of assigning a confidence level to $\boldsymbol{\mu} \in \Re_1$: $\quad \hat{\boldsymbol{\mu}} \sim \mathrm{F} = \mathrm{N}_2(\boldsymbol{\mu}, \mathrm{I})$

$$\hat{\boldsymbol{\mu}}^{**} \sim \hat{\mathrm{F}} = \mathrm{N}_2(\hat{\boldsymbol{\mu}}_0, \mathrm{I})$$

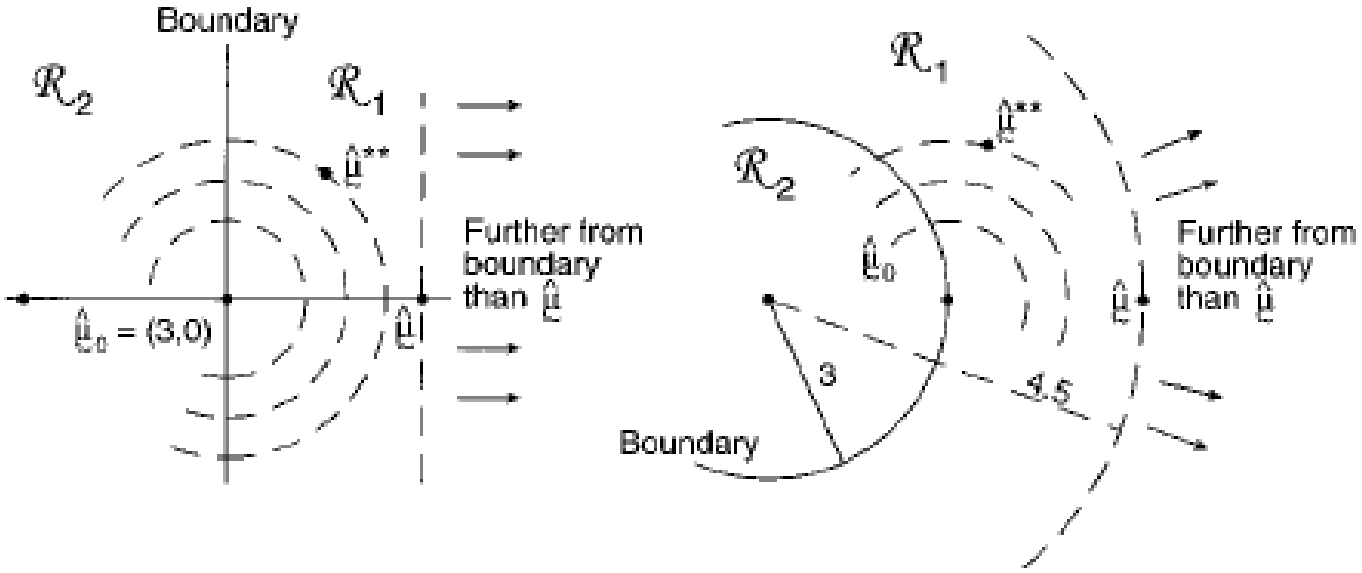


FIG. 5. Confidence levels of the two cases in Fig. 4; $\hat{\boldsymbol{\mu}}_0 = (3, 0)$ is the closest point to $\hat{\boldsymbol{\mu}} = (4.5, 0)$ on the boundary separating $\mathcal{R}_1$ from $\mathcal{R}_2$; bootstrap vector $\hat{\boldsymbol{\mu}}^{**} \sim N_2(\hat{\boldsymbol{\mu}}_0, I)$. The confidence level $\hat{\alpha}$ is the probability that $\hat{\boldsymbol{\mu}}^{**}$ is closer than $\hat{\boldsymbol{\mu}}$ to the boundary.

- p-value = $1-\hat{\alpha}$ =

$$\frac{\#\left\{\hat{\boldsymbol{\mu}}^{**} \text{ further than } \hat{\boldsymbol{\mu}} \text{ from (the boundary of) } \Re_2\right\}}{B}$$

- The confidence level for the two cases (computed numerically):

$$\hat{\alpha}_{\mathrm{I}} = 0.933 \ \left(= \tilde{\alpha}_{\mathrm{I}}, \ \text{Felsenstein's confidence level}\right)$$

$$\hat{\alpha}_{\mathrm{II}} = 0.914 \ \left(< \tilde{\alpha}_{\mathrm{II}}, \ \text{Felsenstein's confidence level}\right)$$

Why are the answers different?

- The boundary curves *away* from $\hat{\mu}$ $\Rightarrow$ The probabilistic distance from $\hat{\mu}$ to $\Re_2$ $(\tilde{\alpha})$ > the probabilistic distance from $\hat{\mu}_0$ to $\hat{\mu}$ $(\hat{\alpha})$.

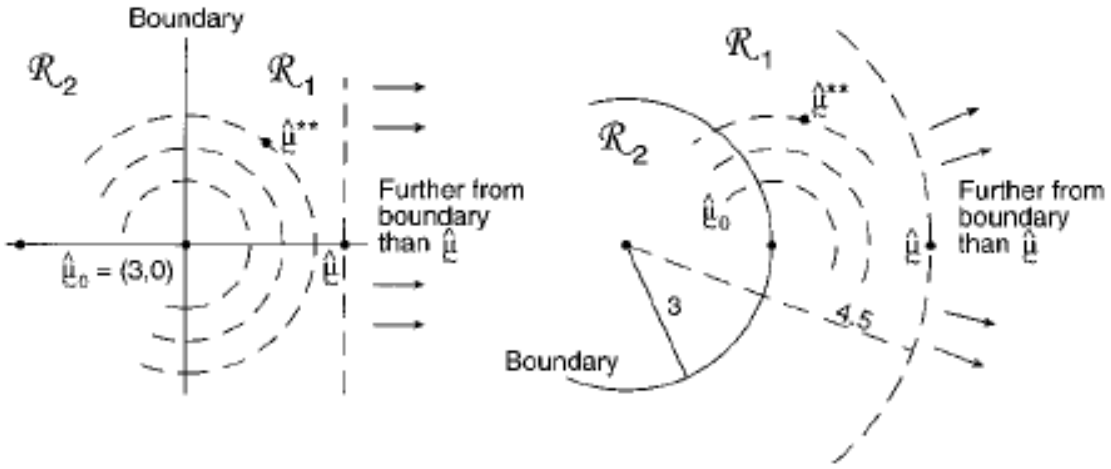


FIG. 5. Confidence levels of the two cases in Fig. 4; $\hat{\mu}_0 = (3, 0)$ is the closest point to $\hat{\mu} = (4.5, 0)$ on the boundary separating $\mathcal{R}_1$ from $\mathcal{R}_2$; bootstrap vector $\hat{\mu}^{**} \sim N_2(\hat{\mu}_0, I)$. The confidence level $\hat{\alpha}$ is the probability that $\hat{\mu}^{**}$ is closer than $\hat{\mu}$ to the boundary.

$\Rightarrow$ $\tilde{\alpha}$ is *not* systematically biased downward.

The “bias” depends on the geometrics of the problem. In this example, Felsenstein’s confidence is biased upward.

c) Efron's method for better assessment of confidence

- Efron suggests an approximation formula for converting Felsenstein's confidence level $\tilde{\alpha}$ to a hypothesis-testing confidence level $\hat{\alpha}$.
- Denote: $\tilde{z} = \Phi^{-1}(\tilde{\alpha}), \quad \hat{z} = \Phi^{-1}(\hat{\alpha})$

$$z_0 = \Phi^{-1}\left(\mathrm{P}_{\hat{\boldsymbol{\mu}}_0}\{\hat{\boldsymbol{\mu}}^{**} \epsilon \Re_1\}\right)$$

- Then in the normal case: $\hat{z} \doteq \tilde{z} - 2z_0$

- z_0 is of order $\frac{1}{\sqrt{n}}$, and the error in estimating $\hat{z}$ is of order $\frac{1}{n}$. This is called "second order accuracy".

Estimating $\hat{z}$ in the normal example:

- $\hat{\boldsymbol{\mu}}^* \sim \mathrm{N}_2(\hat{\boldsymbol{\mu}}, \mathrm{I}) \longrightarrow \hat{\boldsymbol{\mu}}^*(1), \cdots, \hat{\boldsymbol{\mu}}^*(B)$

$$\tilde{z} = \Phi^{-1}\left(\frac{\#\left\{\hat{\boldsymbol{\mu}}^* \text{ vectors in } \Re_1\right\}}{B}\right), \quad B \sim 100$$

- $\hat{\boldsymbol{\mu}}^{**} \sim \mathrm{N}_2(\hat{\boldsymbol{\mu}}_0, \mathrm{I}) \longrightarrow \hat{\boldsymbol{\mu}}^{**}(1), \cdots, \hat{\boldsymbol{\mu}}^{**}(B_2)$

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\hat{\boldsymbol{\mu}}^{**} \text{ vectors in } \Re_1\right\}}{B_2}\right), \quad B_2 \sim 2000$$

- $\hat{z} \doteq \tilde{z} - 2z_0$

- In the multinomial model (trees)

$$\hat{z} \doteq \frac{\tilde{z} - z_0}{1 + a \cdot (\tilde{z} - z_0)} - z_0$$

- a - *acceleration* factor from BC_a

Example with malaria data:

- Efron got $\tilde{\alpha} = 0.965$ for monophyly of 9-10 clade with $B = 200$. With $B = 2000$ he got $\tilde{\alpha} = 0.962$.
- He wants to compute $\hat{\alpha}$.

- Denote: $P^{(\text{cent})} = \left(\frac{1}{n}, \cdots, \frac{1}{n}\right)$

$$P^* = (P_1^*, \cdots, P_n^*) \sim \text{Mult}\left(P^{(\text{cent})}\right)$$

- P^* - vector of proportions of the original matrix' columns in a bootstrap sample

- A bootstrap sample (X^* matrix): $P^* \to \hat{D}^* \to \text{trêe}^*$
- The original data (X matrix): $P \to \hat{D} \to \text{trêe}$

Computing $\hat{\alpha}$:

Step 1:

- $P^*(1), \cdots, P^*(B_2) \sim \text{Mult}\left(P^{(\text{cent})}\right), \qquad B = 2000$
- 9-10 clade was monophyletic in 1923 out of 2000 samples.

$$\Rightarrow \quad \tilde{\alpha} = \frac{1923}{2000} = 0.962$$

Step 2:

- Out of the first 200 bootstrap vectors, 9-10 was *not* monophyletic in 7: $P^{(1)}, \ldots, P^{(7)}$
- For each, choose $0 \le w \le 1$ s.t. the vector $p^{(j)} = wP^{(j)} + (1-w)P^{(cent)}$ is on the boundary of 9-10 monophyly. Use binary search to do it.
- The vectors $p^{(j)}$ play the role of $\hat{\mu}_0$.

Step 3:

- For each boundary vector $p^{(j)}$, create $B_2 = 400$ bootstrap samples: $P^{**}(1), \cdots, P^{**}(B_2) \sim \text{Mult}\left(p^{(j)}\right)$
- Each P^{**} gives a tree. The trees where 9-10 is monophyletic were counted:

Case	No.	B_2
1	218	400
2	204	400
3	223	400
4	214	400
5	213	400
6	216	400
7	223	400
Total	1511	2800

$$\Rightarrow \quad \hat{z}_0 = \Phi^{-1}\left(\frac{1511}{2800}\right) = 0.0995$$

Step 4:

- Given a direction vector $U = p^{(j)} - P^{(cent)}$ from the center $(\hat{\mu}, P^{(cent)})$ to the boundary $(\hat{\mu}_0, p^{(j)})$,

$$a(U) = \frac{\frac{1}{6}\sum_{k=1}^{n} U_k^3}{\left(\sum_{k=1}^{n} U_k^2\right)^{3/2}}$$

- a values they got:

Case	a
1	0.014
2	0.009
3	0.014
4	0.012
5	0.014
6	0.012
7	0.014
Average	0.0129

- The average: $a = 0.0129$

Step 5:

- $\tilde{z} = \Phi^{-1}(\tilde{\alpha}) = \Phi^{-1}(0.962) = 1.77$

  $\hat{z}_0 = 0.0995$

  $a = 0.0129$

$$\Rightarrow \quad \hat{z} \doteq \frac{\tilde{z} - \hat{z}_0}{1 + a \cdot (\tilde{z} - \hat{z}_0)} - \hat{z}_0 = 1.5357$$

$$\Rightarrow \quad \hat{\alpha} = \Phi(\hat{z}) = 0.938 \; < \; 0.962 = \tilde{\alpha}$$

- In this example Felsenstein's $\tilde{\alpha}$ was biased *upward*. This happens when $z_0 > 0$.
- The opposite can also happen, for example with the 7-8 clade.

Summary

- In a Bayesian sense, $\tilde{\alpha}$ is a reasonable assessment of confidence.
- $\tilde{\alpha}$ is *not* systematically biased downward.
- $\hat{\alpha}$ has a more familiar interpretation as hypothesis-testing confidence level.
- $\hat{\alpha}$ can be estimated by a two-level bootstrap algorithm.
- For $\tilde{\alpha}$, 50-100 bootstrap samples at the first level are enough. For $\hat{\alpha}$, ~2000 bootstrap samples are needed at the second level.