

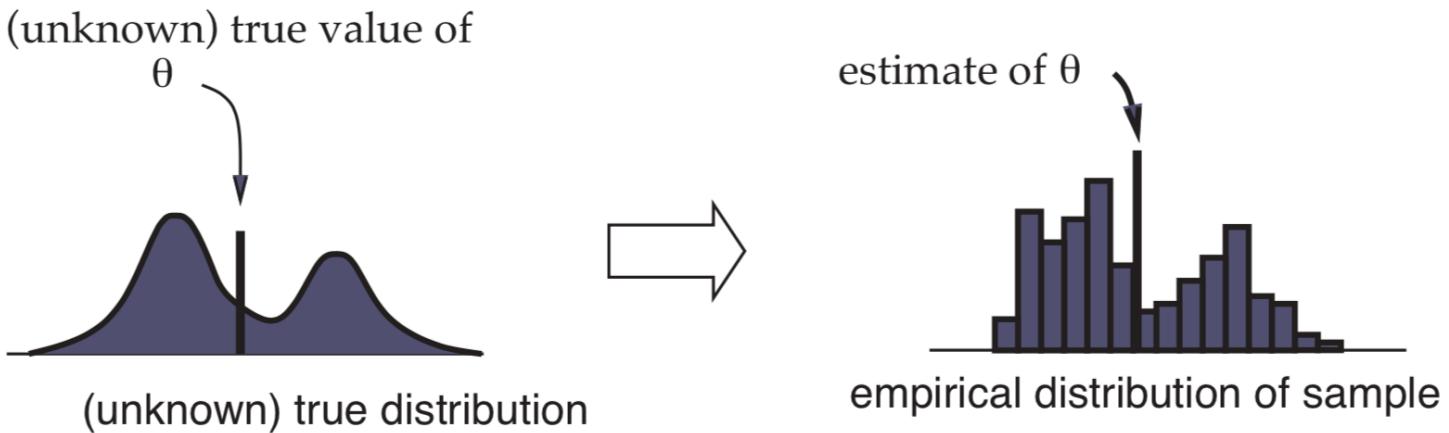
# Testing phylogenetic hypotheses

Slides courtesy of Mark Holder and Joe Felsenstein

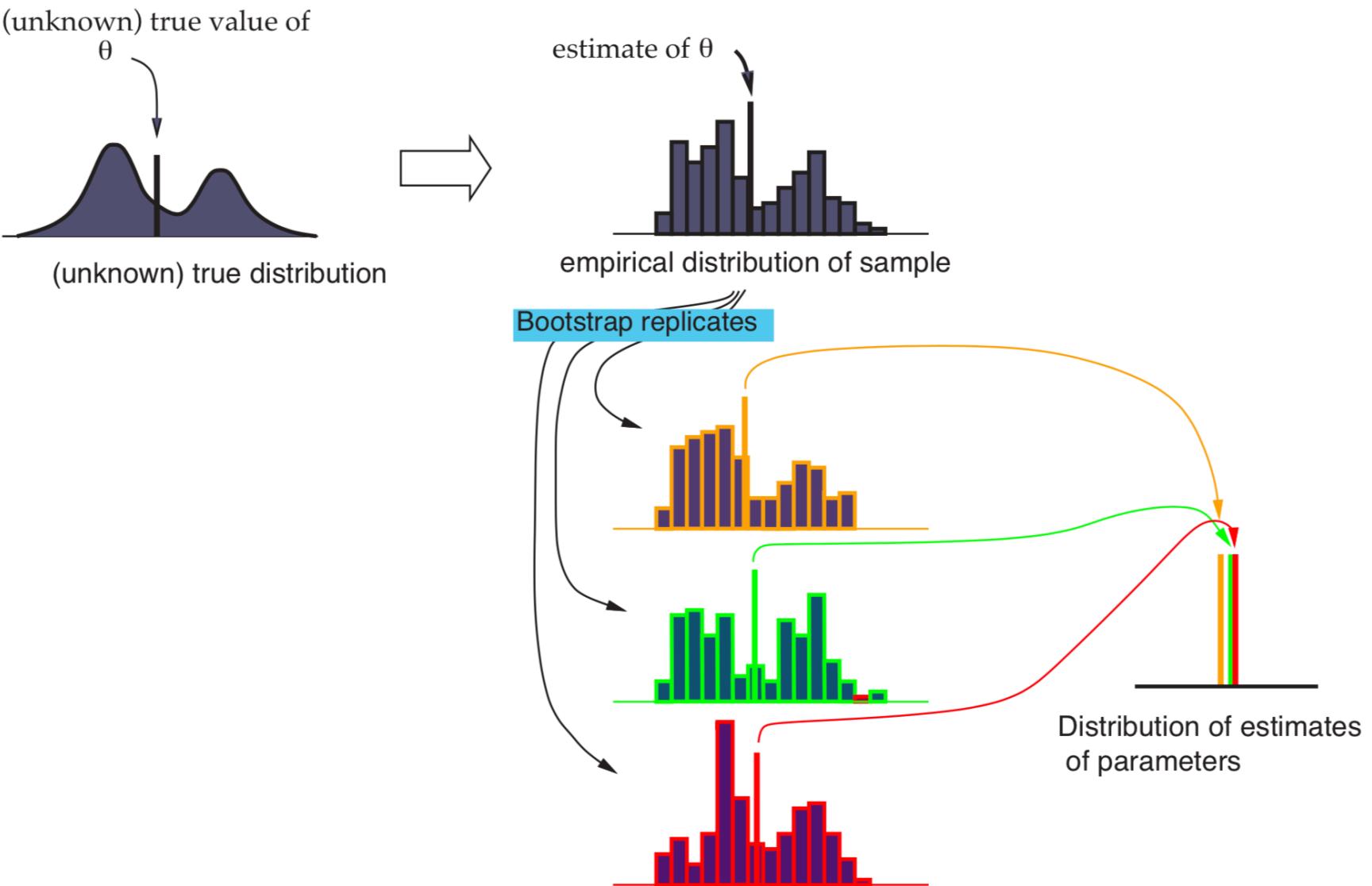
Download the nexus file for this weeks lab and execute it in MrBayes...

```
$ mb RootOfBilateria_Constrained-Bayes.nxs
```

# The bootstrap

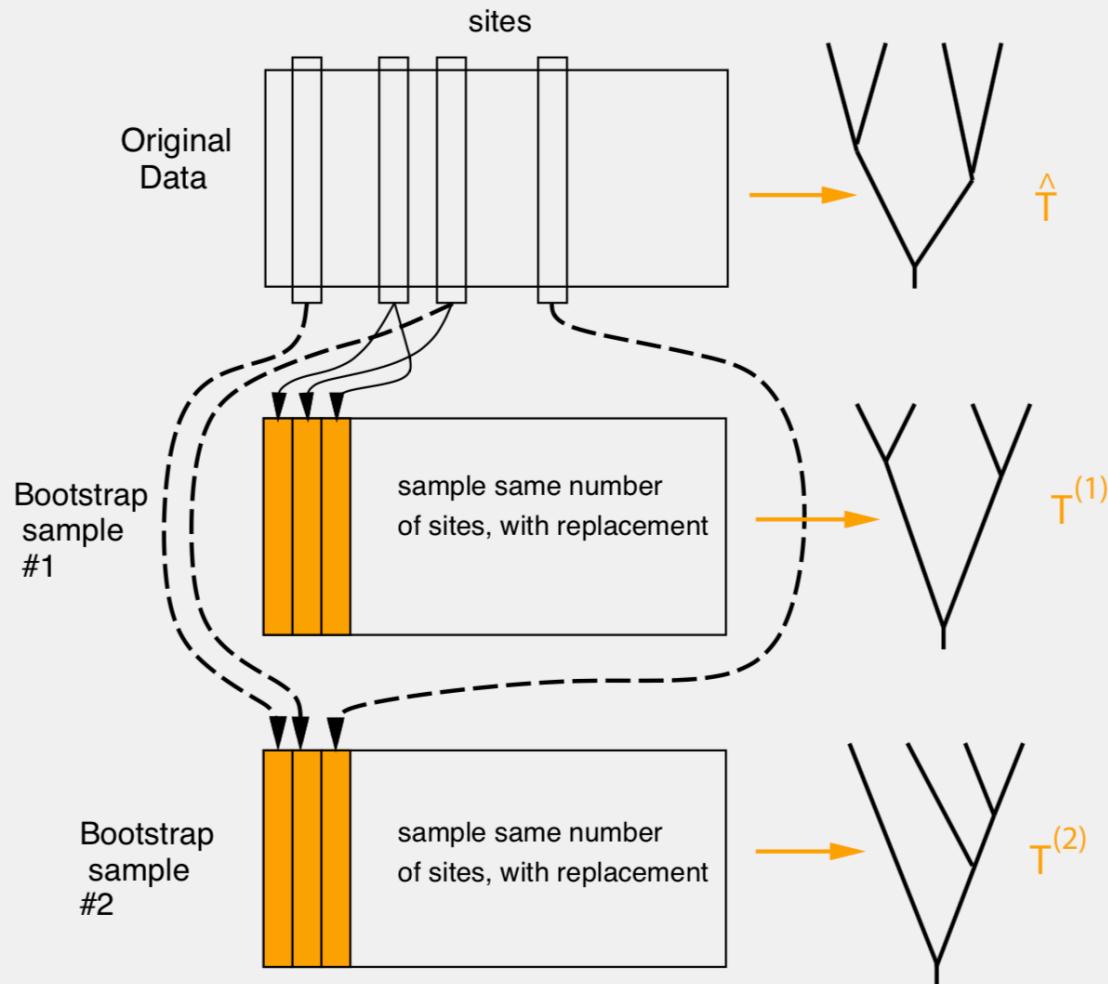


# The bootstrap



Slide from Joe Felsenstein

# The bootstrap for phylogenies

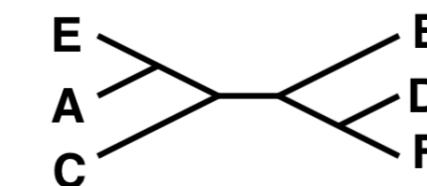
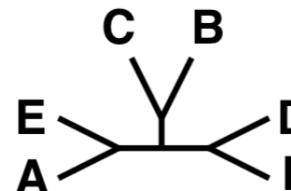
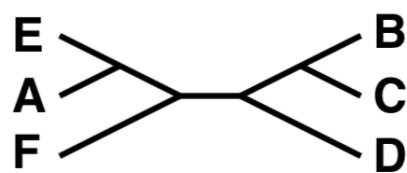
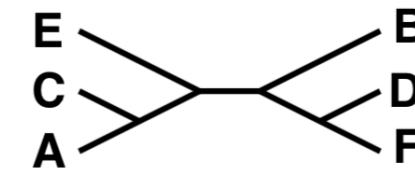
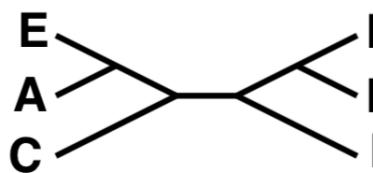


Slide from Joe Felsenstein

(and so on)

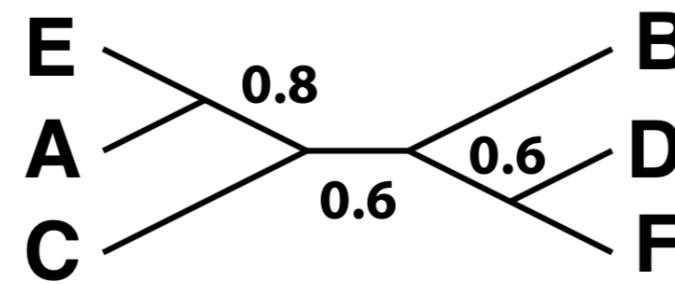
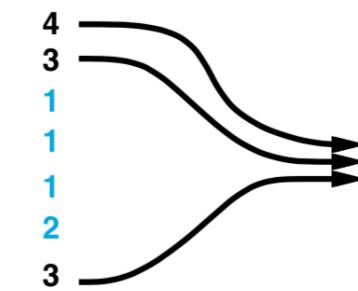
# The majority-rule consensus tree

Trees:

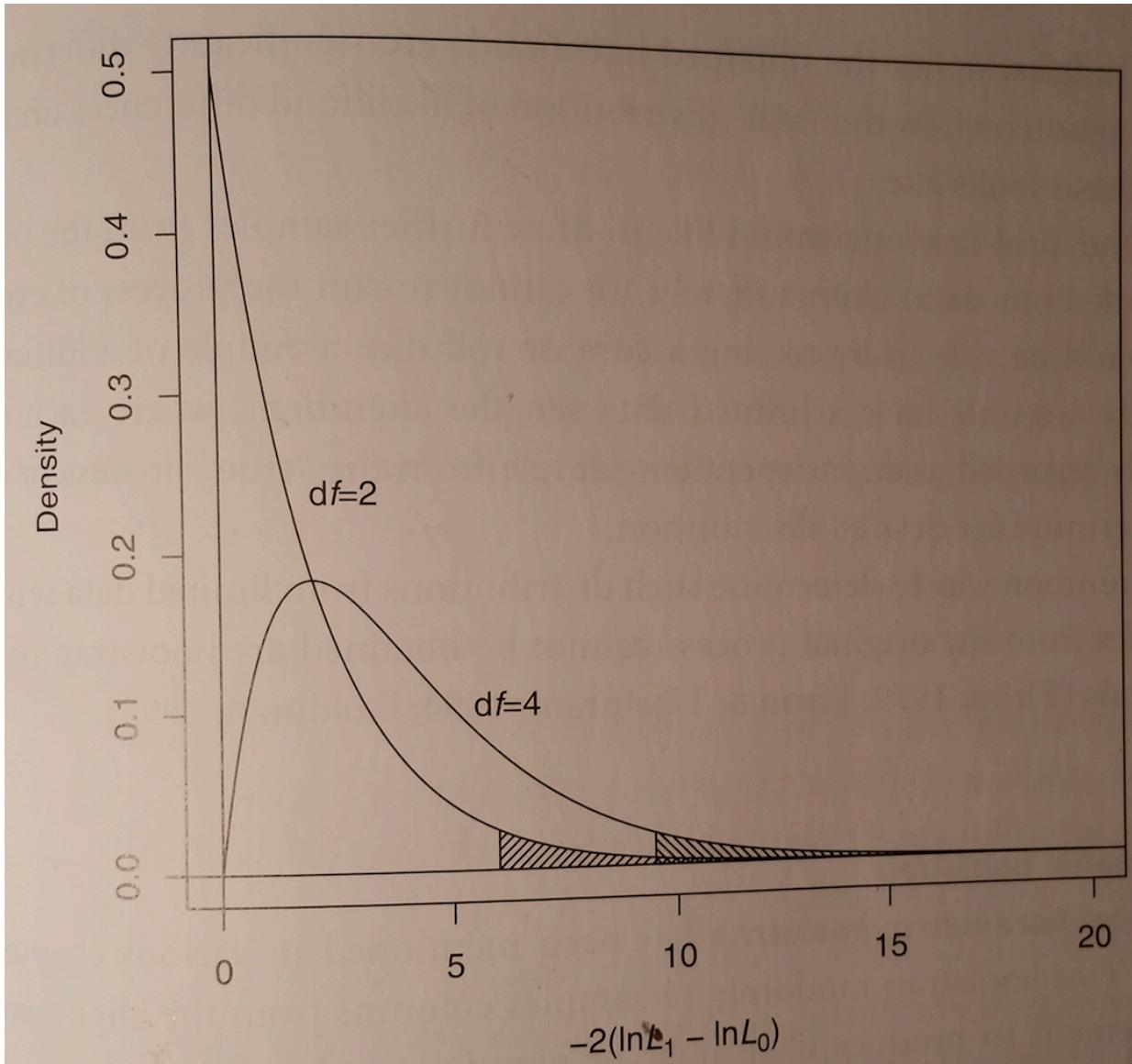


How many times each partition of species is found:

AE   BCDF	4
ACE   BDF	3
ACEF   BD	1
AC   BDEF	1
AEF   BCD	1
ADEF   BC	2
ABCE   DF	3



# Comparing phylogenetic trees using log likelihood ratios



We could use log likelihood ratios to test for fit of substitution model since the models are nested and the test statistic is approx. chi2 distributed.

However, trees are not nested in each other!

## **The simplest phylogenetic test would compare two trees**

Null: If we had no sampling error (infinite data)  $T_1$  and  $T_2$  would explain the data equally well.

Test Statistic:

$$\delta(T_1, T_2 | X) = 2 [\ln L(T_1 | X) - \ln L(T_2 | X)]$$

Expectation under null:

$$\mathbb{E}_{H_0} [\delta(T_1, T_2 | X)] = 0$$

## Frequentist hypothesis testing: coin flipping example

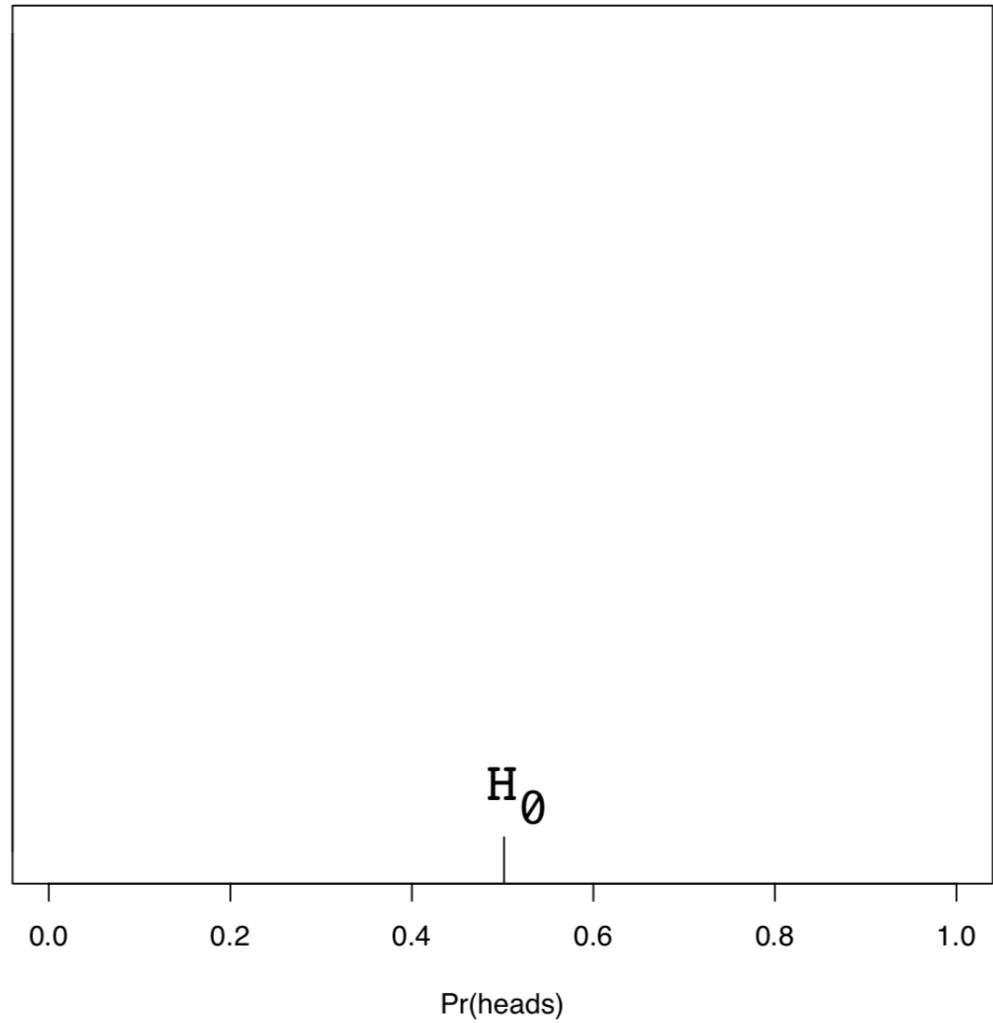
---

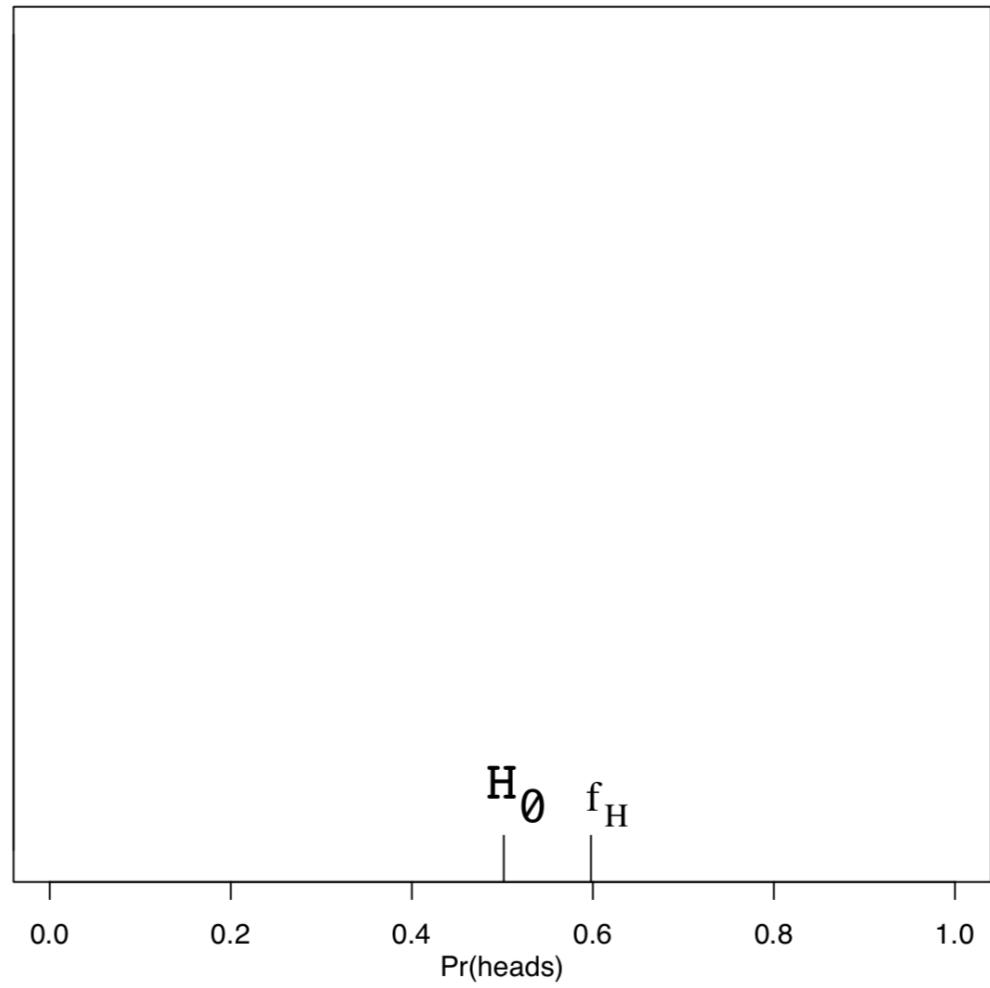
$N = 100$  and  $h = 60$

Can we reject the fair coin hypothesis?  $H_0 : \Pr(\text{heads}) = 0.5$

The “recipe” is:

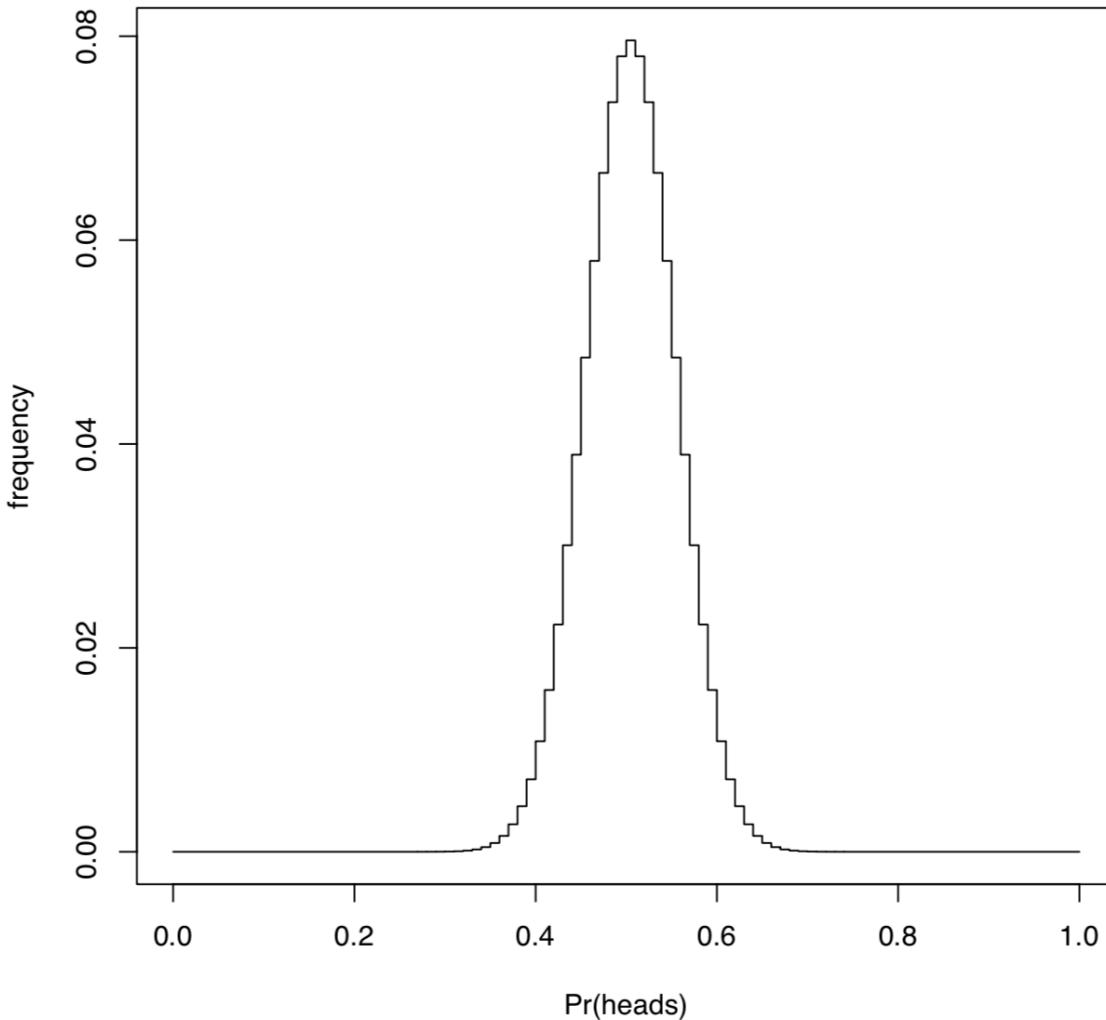
1. Formulate null ( $H_0$ ) and alternative ( $H_A$ ) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic:  $f_H =$  fraction of heads in sample.  
 $f_H = 0.6$
4. Characterize the null distribution of the test statistic
5. Calculate the  $P$ -value: The probability of a test statistic value more extreme than  $f_H$  arising even if  $H_0$  is true.
6. Reject  $H_0$  if  $P$ -value is  $\leq$  your Type I error rate.

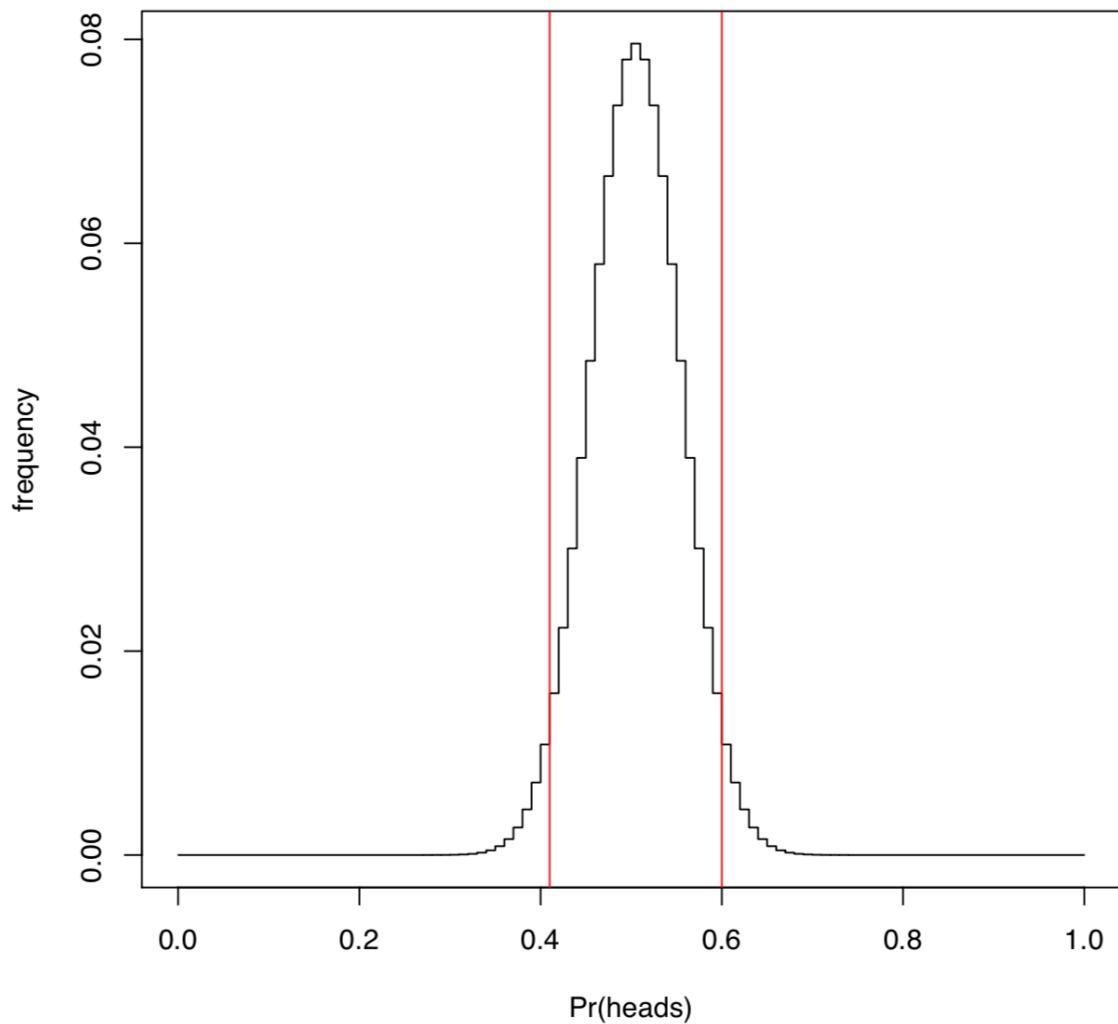




## Null distribution

---





$P\text{-value} \approx 0.058$

Making similar plots for tree inference is hard.

- Our parameter space is trees and branch lengths.
- Our data is a matrix of characters.
- It is hard to put these objects on the same plot.

## **The simplest phylogenetic test would compare two trees**

Null: If we had no sampling error (infinite data)  $T_1$  and  $T_2$  would explain the data equally well.

Test Statistic:

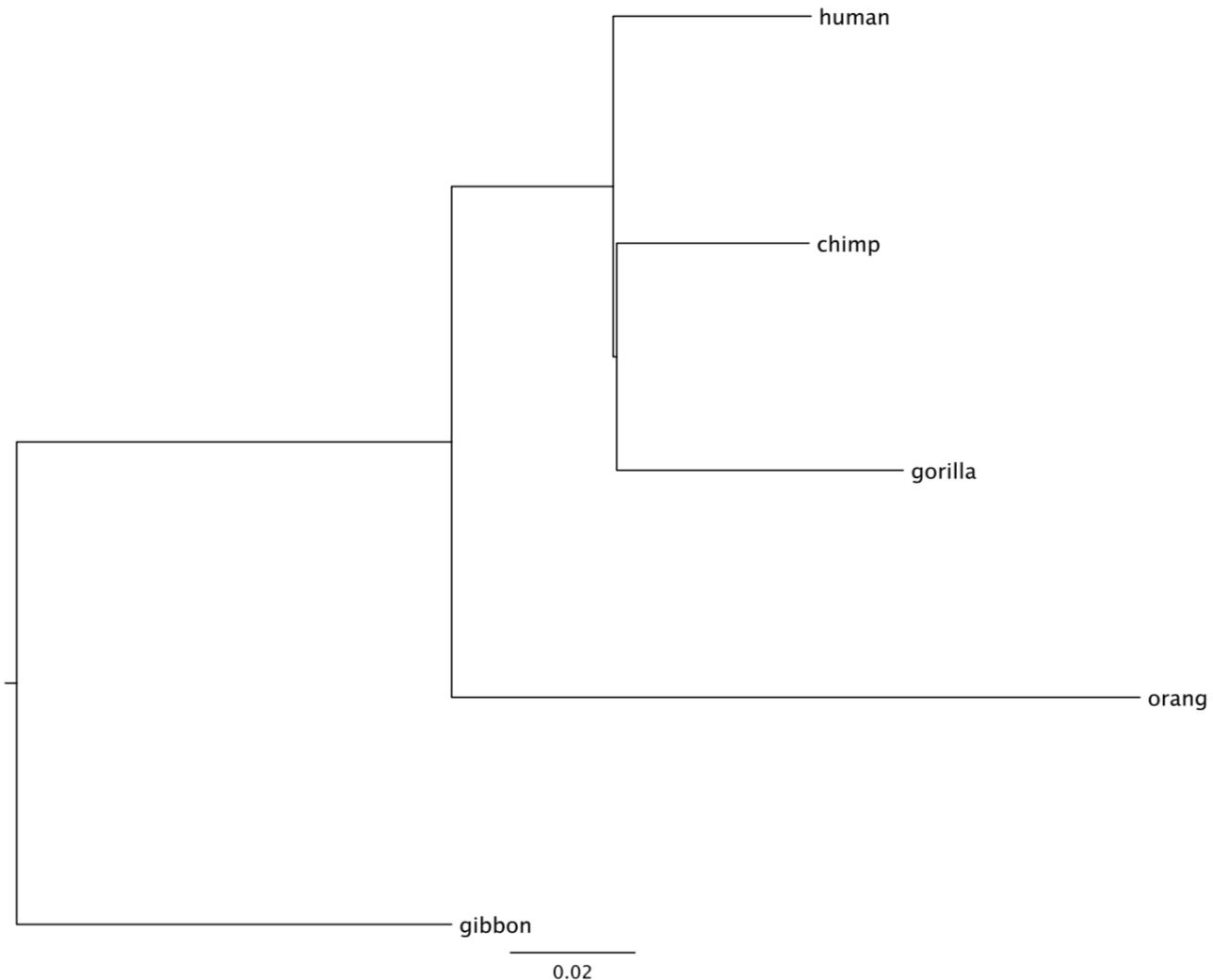
$$\delta(T_1, T_2 | X) = 2 [\ln L(T_1 | X) - \ln L(T_2 | X)]$$

Expectation under null:

$$\mathbb{E}_{H_0} [\delta(T_1, T_2 | X)] = 0$$

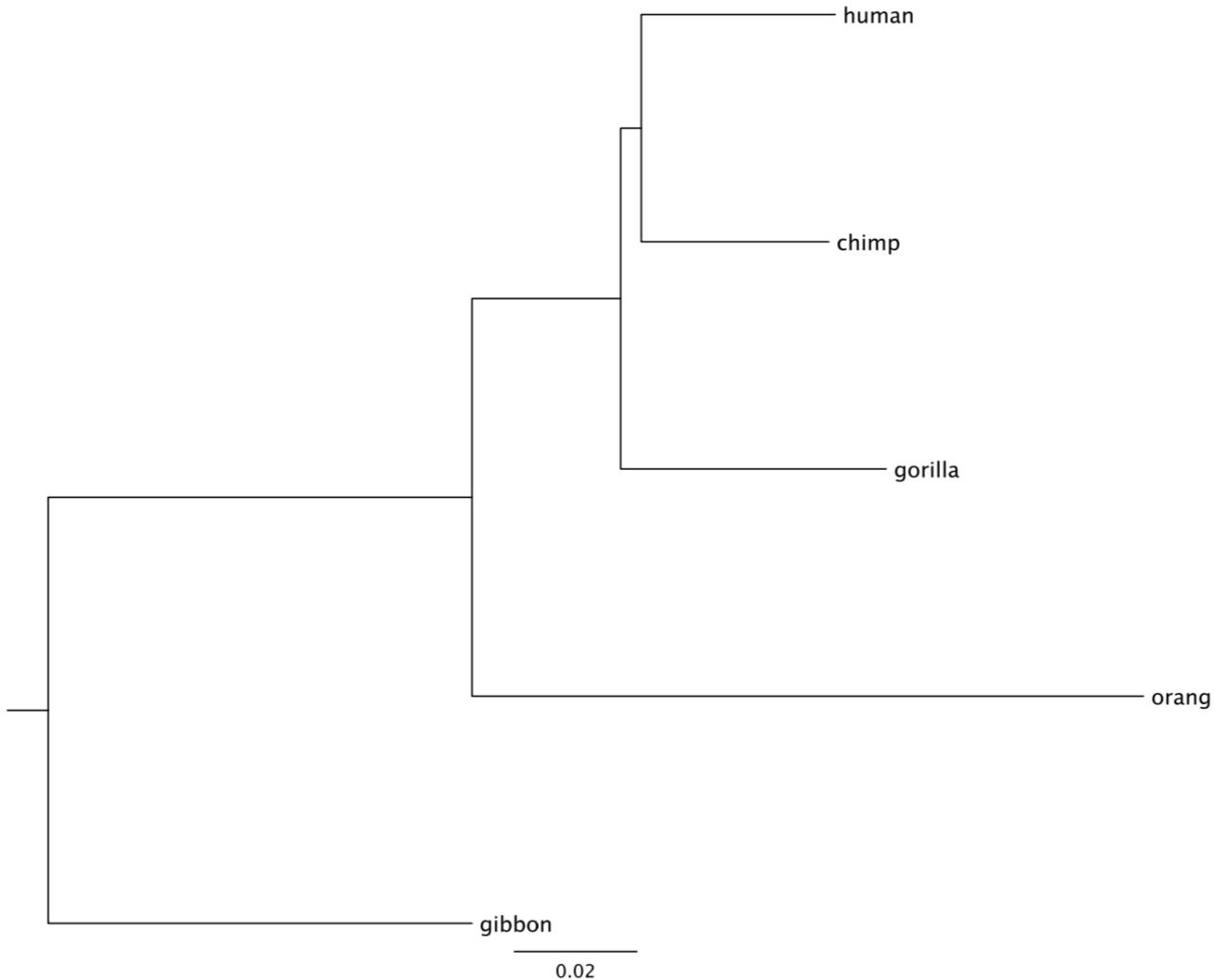
Using 3000 sites of mtDNA sequence for 5 primates

$T_1$  is ((chimp, gorilla), human)



Using 3000 sites of mtDNA sequence for 5 primates

$T_2$  is ((chimp, human), gorilla)



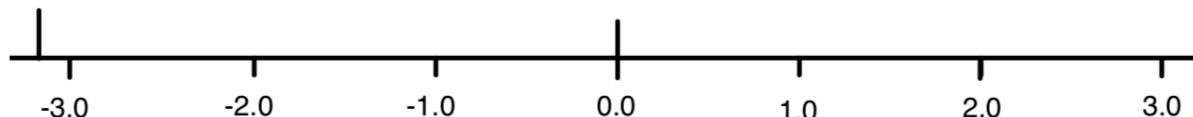
Using 3000 sites of mtDNA sequence for 5 primates

$$T_1 \text{ is } ((\text{chimp}, \text{gorilla}), \text{human}) \quad \ln L(T_1 | X) = -7363.296$$

$$T_2 \text{ is } ((\text{chimp}, \text{human}), \text{gorilla}) \quad \ln L(T_2 | X) = -7361.707$$

$$\delta(T_1, T_2 | X) = -3.18$$

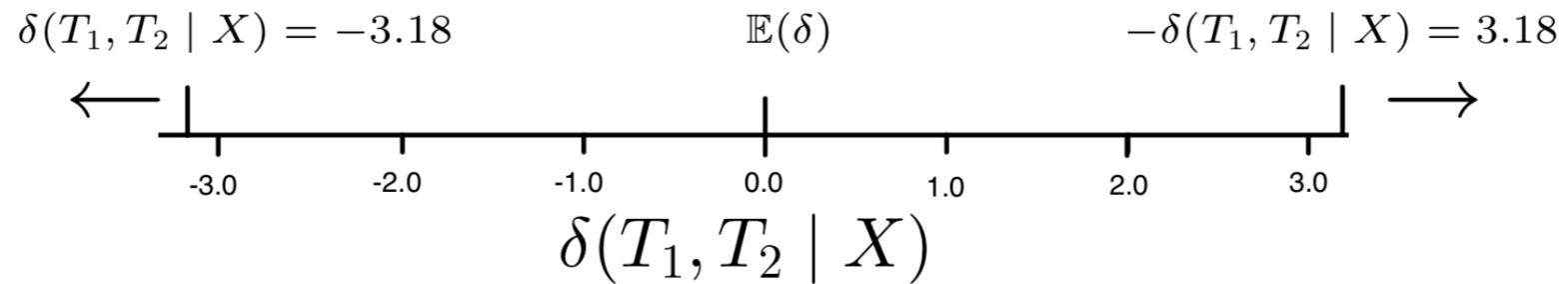
$$\mathbb{E}(\delta)$$



$$\delta(T_1, T_2 | X)$$

To get the  $P$ -value, we need to know the probability:

$$\Pr \left( |\delta(T_1, T_2 | X)| \geq 3.18 \middle| H_0 \text{ is true} \right)$$



## KH Test

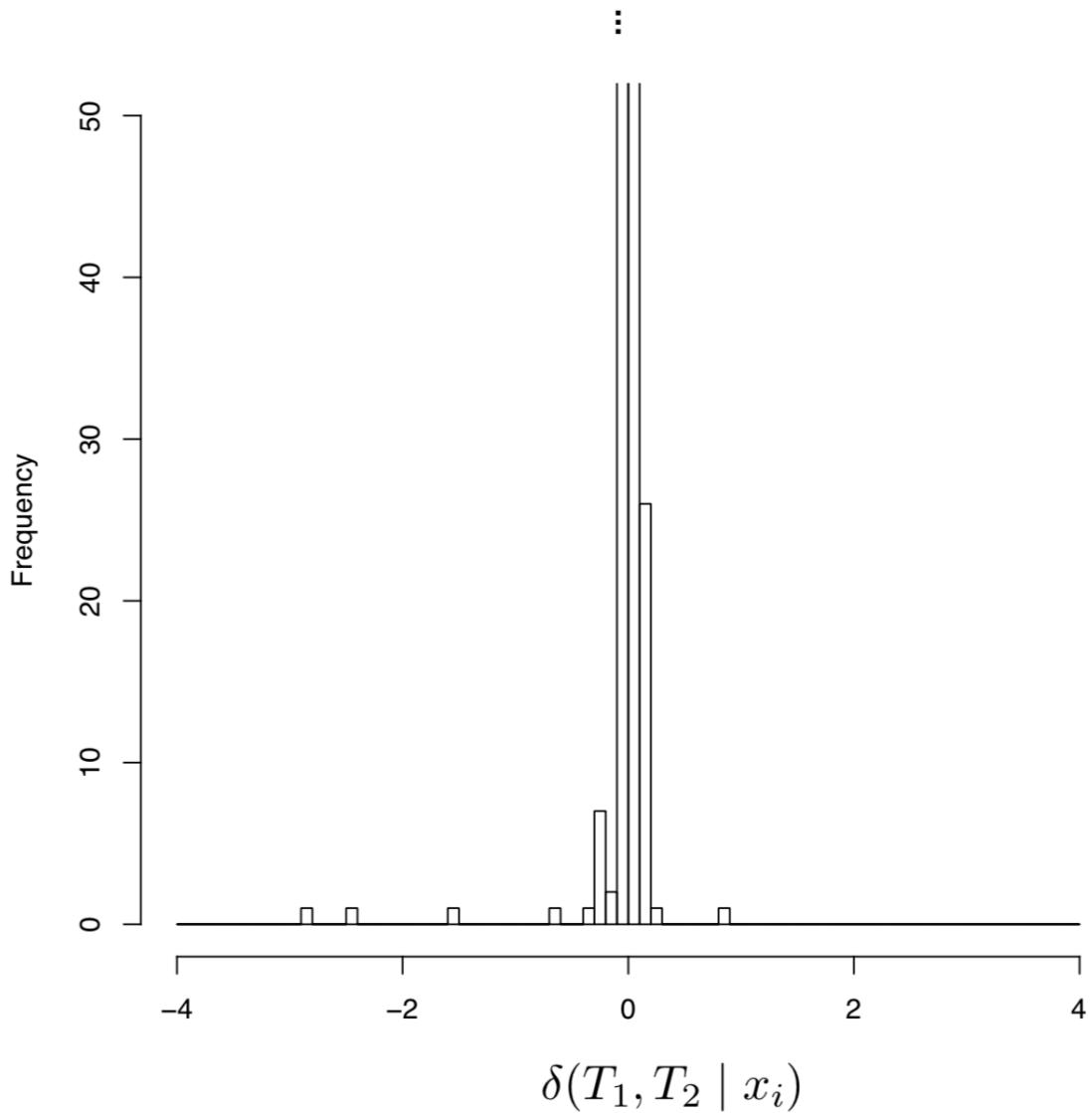
---

1. Examine the difference in  $\ln L$  for each site:  
 $\delta(T_1, T_2 | X_i)$  for site  $i$ .
2. Note that the total difference is simply a sum:

$$\delta(T_1, T_2 | X) = \sum_{i=1}^M \delta(T_1, T_2 | X_i)$$

3. The variance of  $\delta(T_1, T_2 | X)$  will be a function of the variance in “site”  $\delta(T_1, T_2 | X_i)$  values.

$\delta(T_1, T_2 \mid X_i)$  for each site,  $i$ .



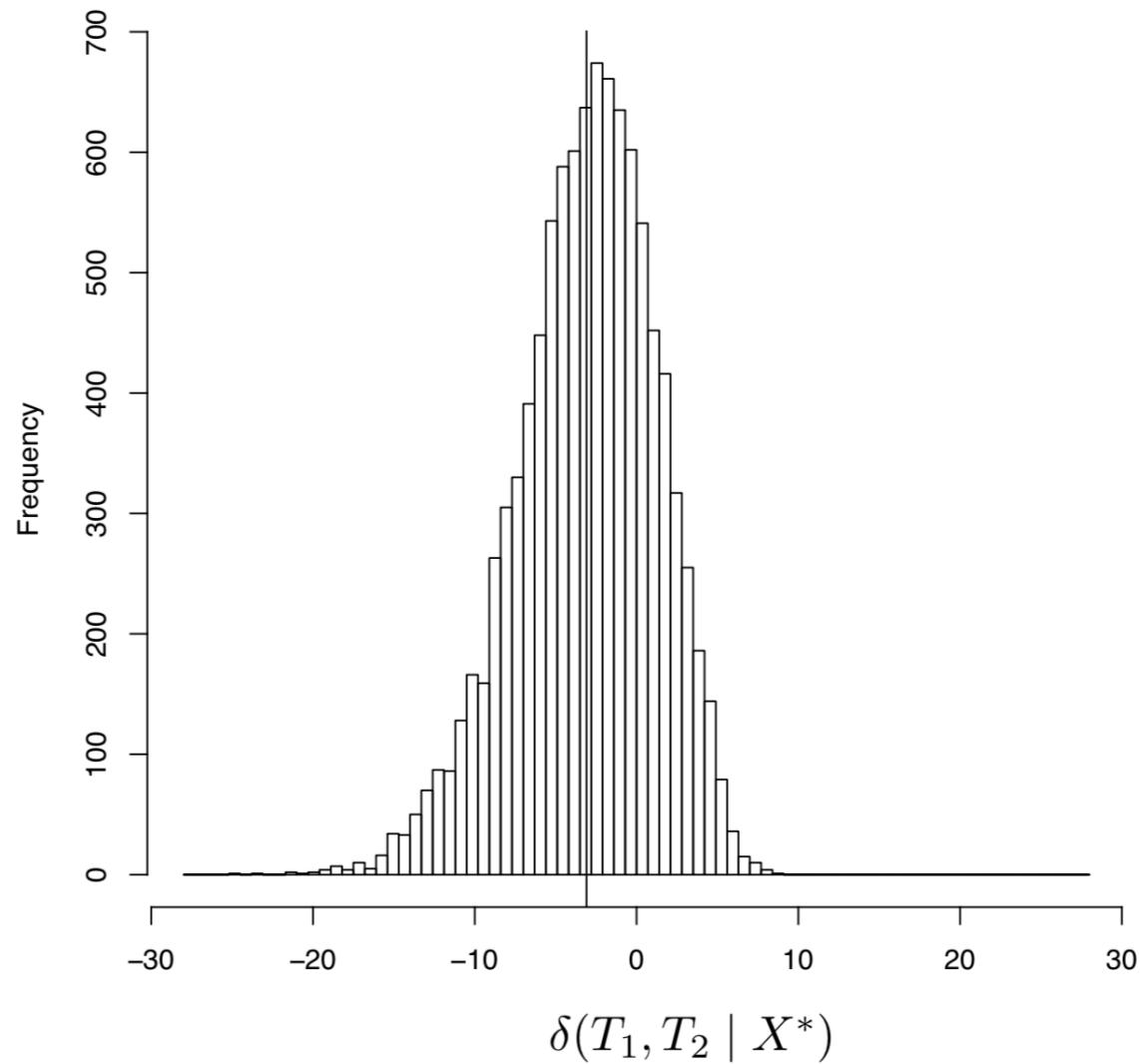
## **KH Test - the variance of $\delta(T_1, T_2 | X)$**

---

To approximate variance of  $\delta(T_1, T_2 | X)$  under the null, we could:

1. use assumptions of Normality (by appealing to the Central Limit Theorem). Or
2. use bootstrapping to generate a cloud of pseudo-replicate  $\delta(T_1, T_2 | X^*)$  values, and look at their variance.

$\delta$  for many (RELL) bootstrapped replicates of the data



## RELL bootstrap

---

Often, the MLE of numerical parameters (including branch lengths) do not change much when we bootstrap.

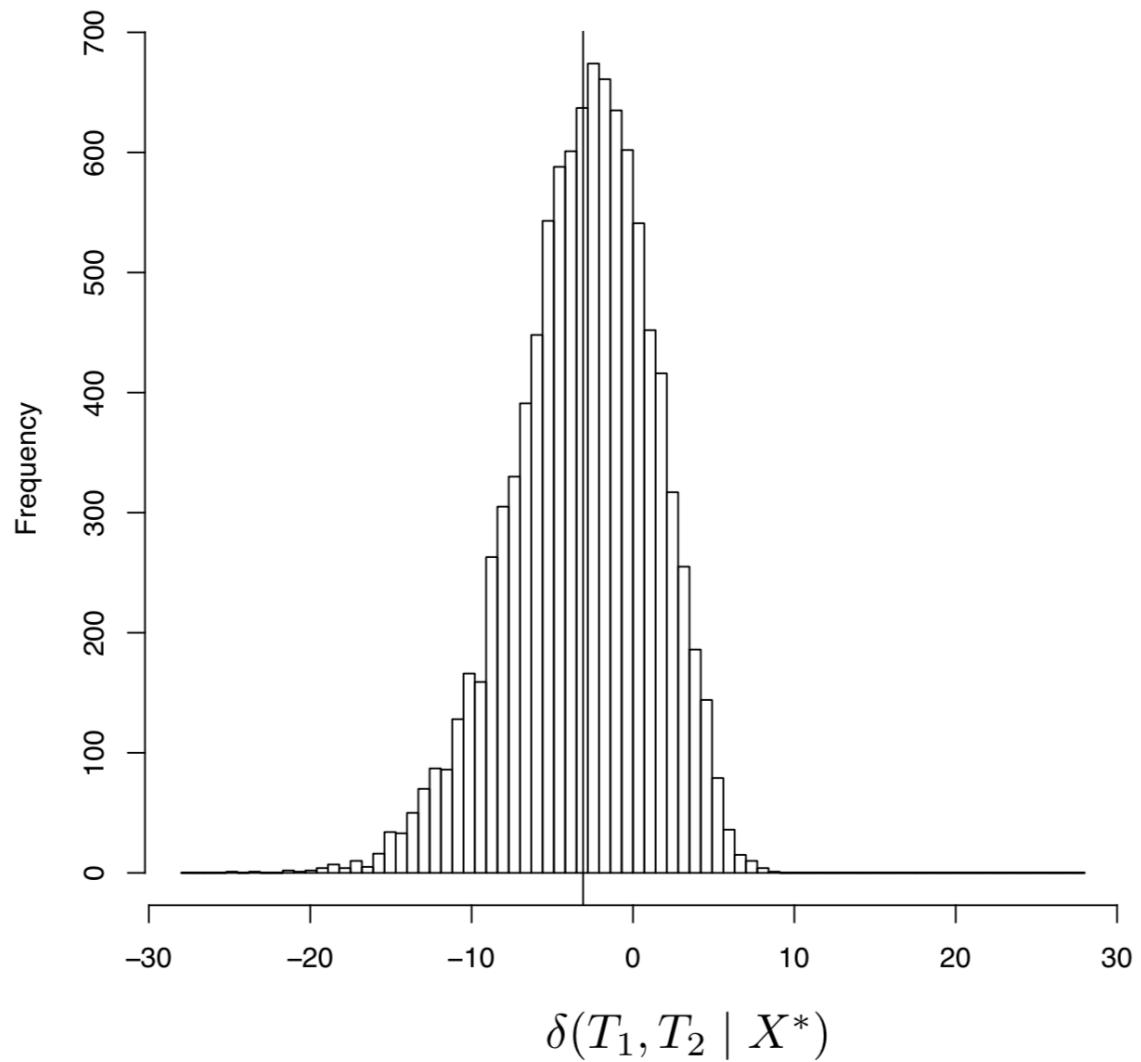
So, we can simply resample the site  $\ln L$  values and sum them (rather than reoptimizing parameters).

This is called the RELL bootstrap (Kishino et al., 1990, and Felsenstein). It is not a “safe” replacement for normal bootstrapping (especially on large trees; Stamatakis et al., 2008) when you want to estimate clade support.

But it should be good enough for helping us learn about the standard error of the  $\ln L$ .

And it is really fast.

The (RELL) bootstrapped sample of statistics.  
Is this the null distribution for our  $\delta$  test statistic?



## KH Test - ‘centering’

---

$H_0$  gives us the expected value:

$$\mathbb{E}_{H_0} [\delta(T_1, T_2 \mid X)] = 0$$

Bootstrapping gives us a reasonable guess of the variance under  $H_0$

By subtracting the mean of the bootstrapped  $\delta(T_1, T_2 \mid X^*)$  values, we can create a null distribution.

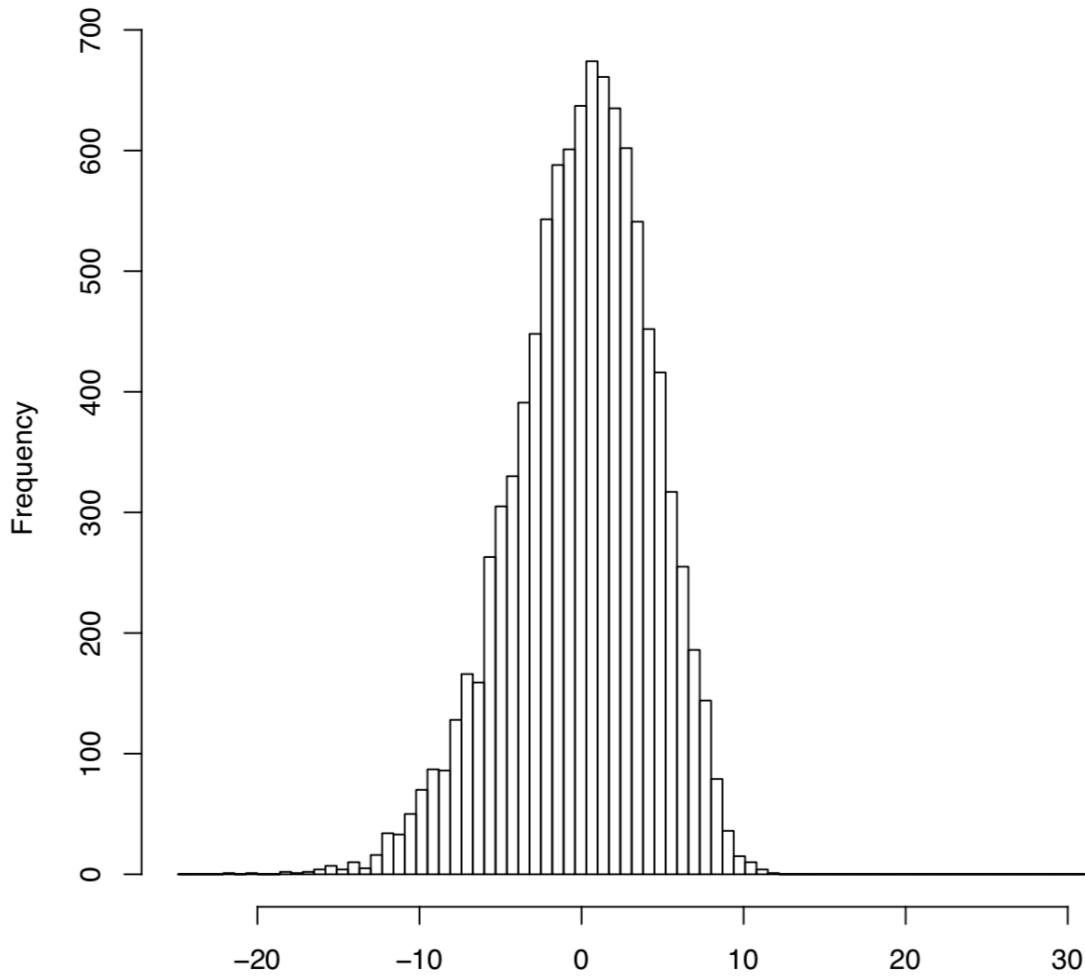
For each of the  $j$  bootstrap replicates, we treat

$$\delta(T_1, T_2 \mid X^{*j}) - \bar{\delta}(T_1, T_2 \mid X^*)$$

as draws from the null distribution.

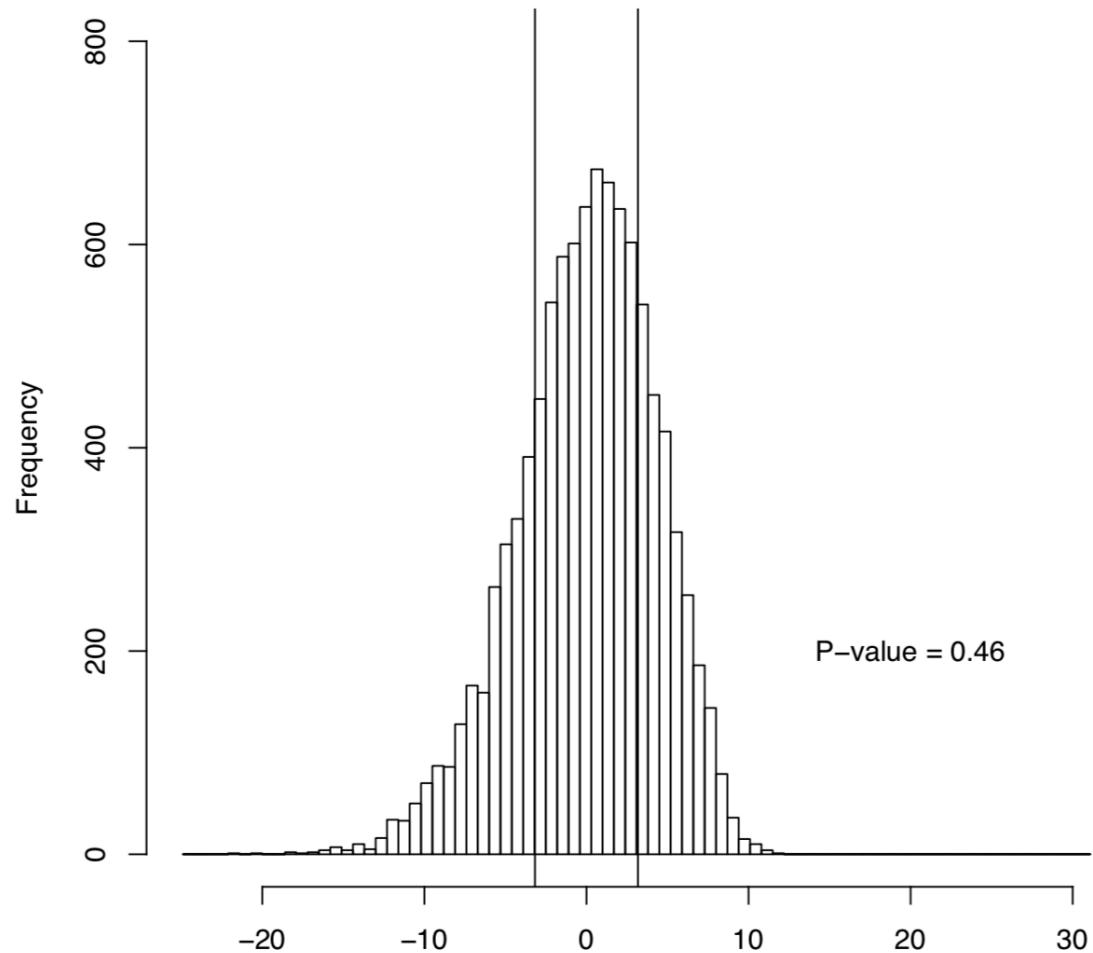
$$\delta(T_1, T_2 \mid X^{(j)}) - \bar{\delta}(T_1, T_2 \mid X^*)$$

for many (RELL) bootstrapped replicates of the data



$$\delta(T_1, T_2 \mid X^{(j)}) - \bar{\delta}(T_1, T_2 \mid X^*)$$

Approximate null distribution with  
tails (absolute value  $\geq 3.18$ ) shown

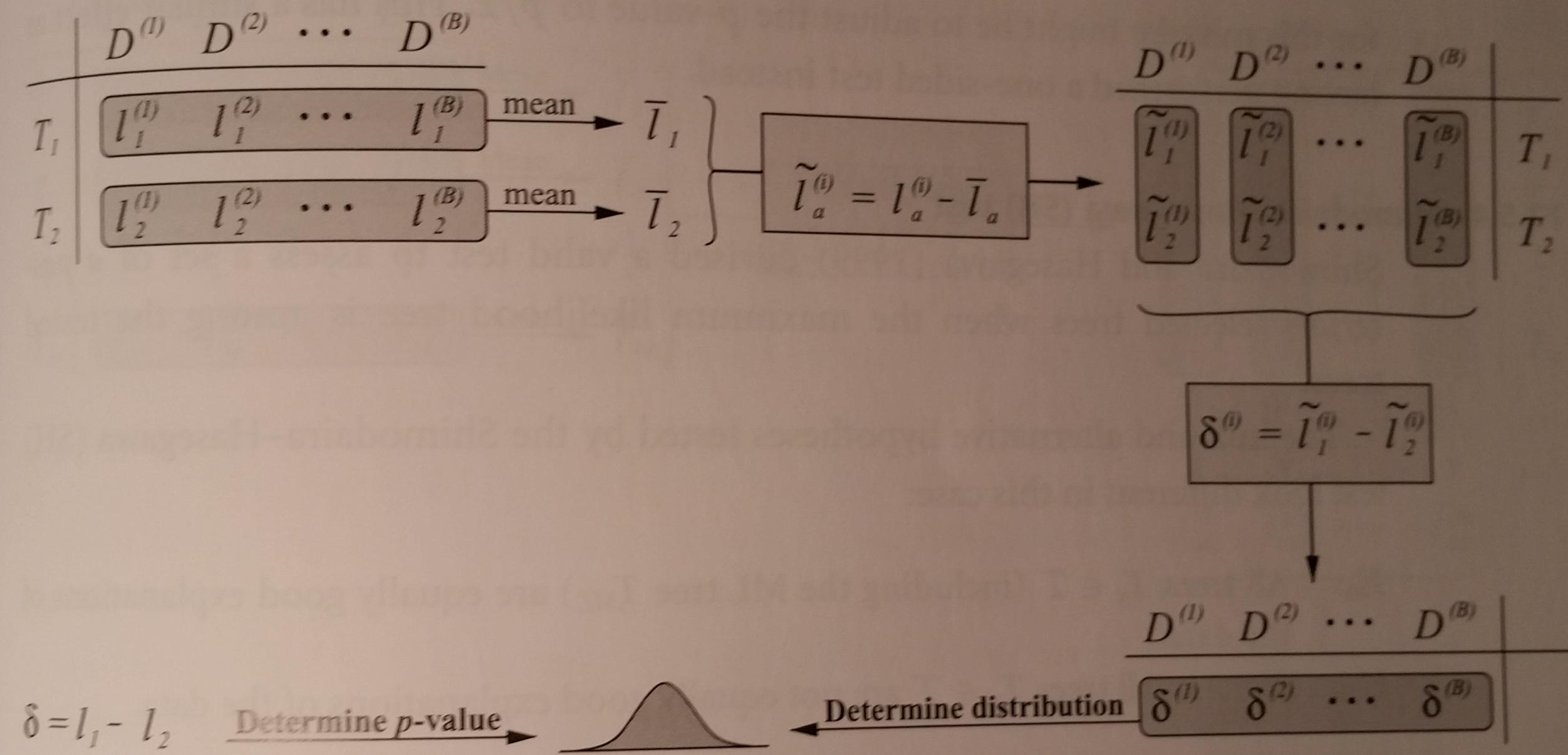


$$\delta(T_1, T_2 | X^*) - \bar{\delta}(T_1, T_2 | X^*)$$

## **Summary - Part 1**

---

- $\delta(T_1, T_2 | X) = 2 [\ln L(T_1 | X) - \ln L(T_2 | X)]$  is a powerful statistic for discrimination between trees.
- We can assess confidence by considering the variance in signal between different characters.
- Bootstrapping helps us assess the variance in  $\ln L$  that we would expect to result from sampling error.



## Scenario

---

1. A (presumably evil) competing lab scoops you by publishing a tree,  $T_1$ , for your favorite group of organisms.
2. You have just collected a new dataset for the group, and your ML estimate of the best tree,  $T_2$ , differ's from  $T_1$ .
3. A KH Test shows that your data **significantly** prefer  $T_2$  over  $T_1$ .
4. You write a (presumably scathing) response article.

Should a *Systematic Biology* publish your response?

**What if start out with only one hypothesized tree, and we want to compare it to the ML tree?**

---

The KH Test is **NOT** appropriate in this context (see Goldman et al., 2000, for discussion of this point)

**Multiple Comparisons:** lots of trees increases the variance of  $\delta(\hat{T}, T_1 | X)$

**Selection bias:** Picking the ML tree to serve as one of the hypotheses invalidates the centering procedure of the KH test.

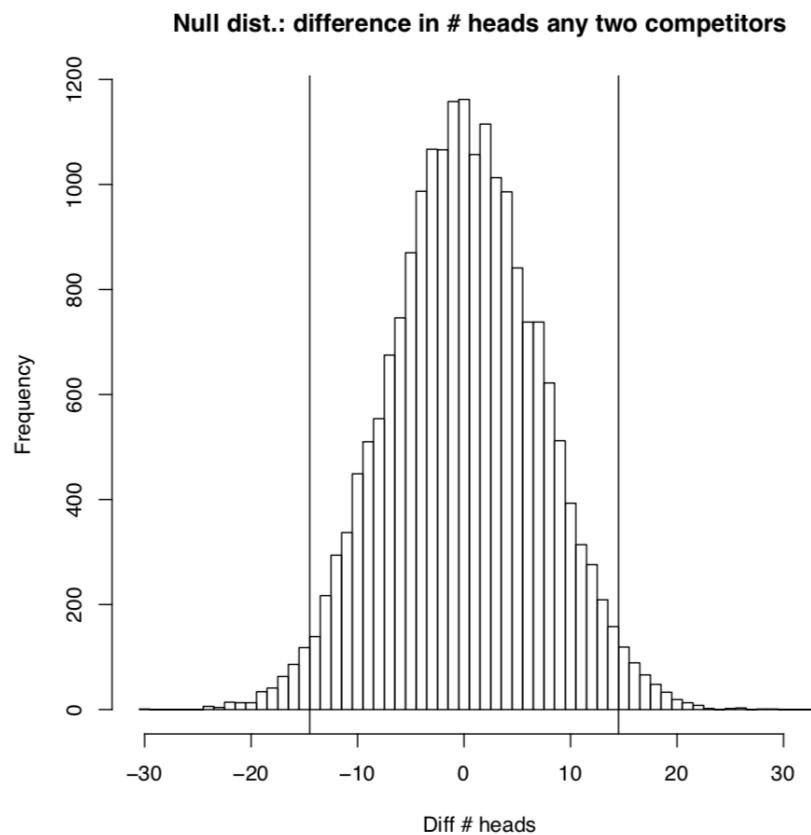
## **Using the ML tree in your test introduces selection bias**

Even when the  $H_0$  is true, we do not expect  
 $2 \left[ \ln L(\hat{T}) - \ln L(T_1) \right] = 0$

Imagine a competition in which a large number of equally skilled people compete, and you compare the score of one competitor against the highest scorer.

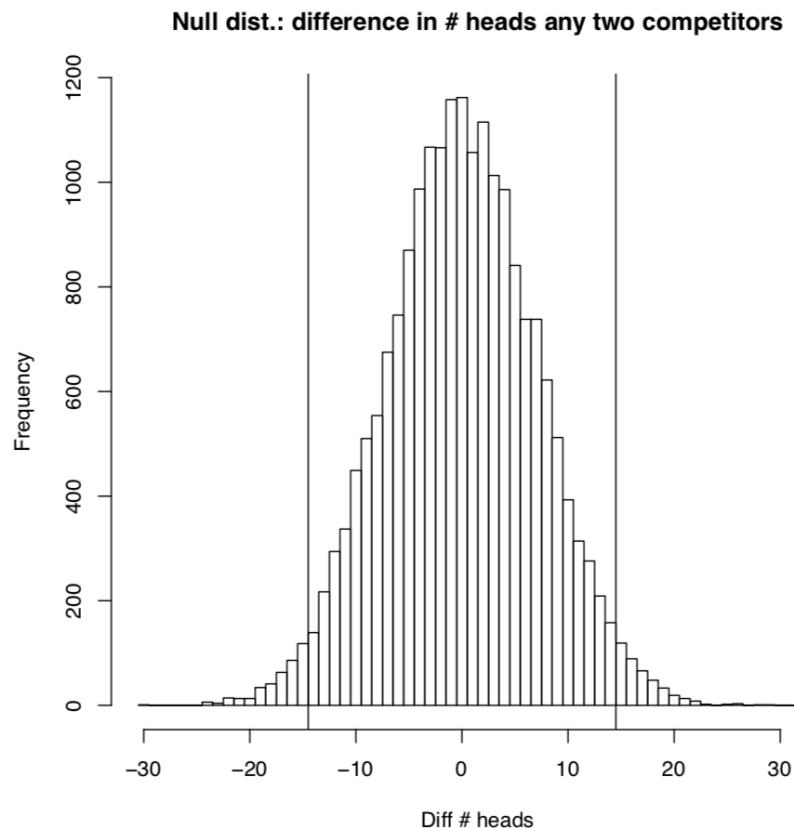
Experiment: 70 people each flip a fair coin 100 times and count # heads.

$$h_1 - h_2$$

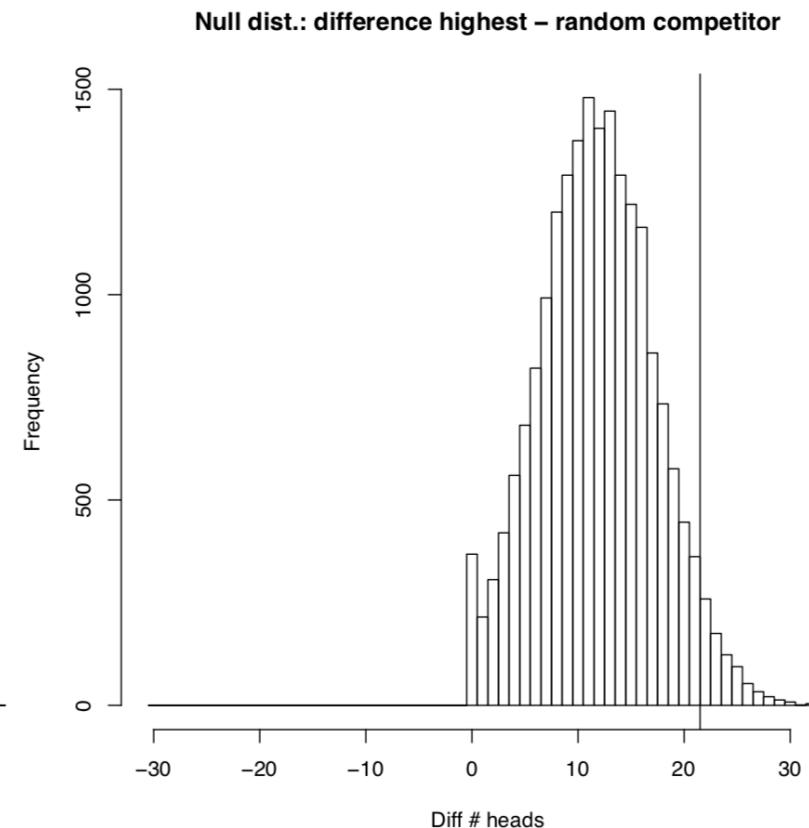


Experiment: 70 people each flip a fair coin 100 times and count # heads.

$$h_1 - h_2$$



$$\max(h) - h_1$$



**Shimodaira and Hasegawa proposed the SH test which deals the “selection bias” introduced by using the ML tree in your test**

---

You have to specify of a **set of candidate trees** - inclusion in this set **must not** depend on the dataset to be analyzed.

The null hypothesis is that all members of the candidate set have the same expected score.

The test makes worst-case assumptions, so the SH test **is conservative**.

## SH test candidate set selection

---

- Should be all trees that you would have seriously entertained before seeing the data (considering a subset of trees for computational convenience can invalidate the test).
- Using all trees is safe.
- If a tree has low  $\ln L$  and low variance of site-log-likelihoods then it can probably be safely removed without affecting the  $P$ -values of other trees<sup>1</sup>

---

<sup>1</sup>Because such a tree would be unlikely to ever be the tree that is the determines the maximum displacement from the centered value,  $m^{(j)}$ .

## SH Test details

- For each tree  $T_i$  in the candidate set calculate  $\delta(\hat{T}, T_i | X)$
- Bootstrap to generate  $\ln L(T_i | X^{(j)})$  for each bootstrap replicate  $j$ .
- For each tree  $T_i$ , use the mean,  $\bar{\ln L}(T_i | X^*)$ , over all bootstrap replicates to center the bootstrapped collection of log-likelihoods:

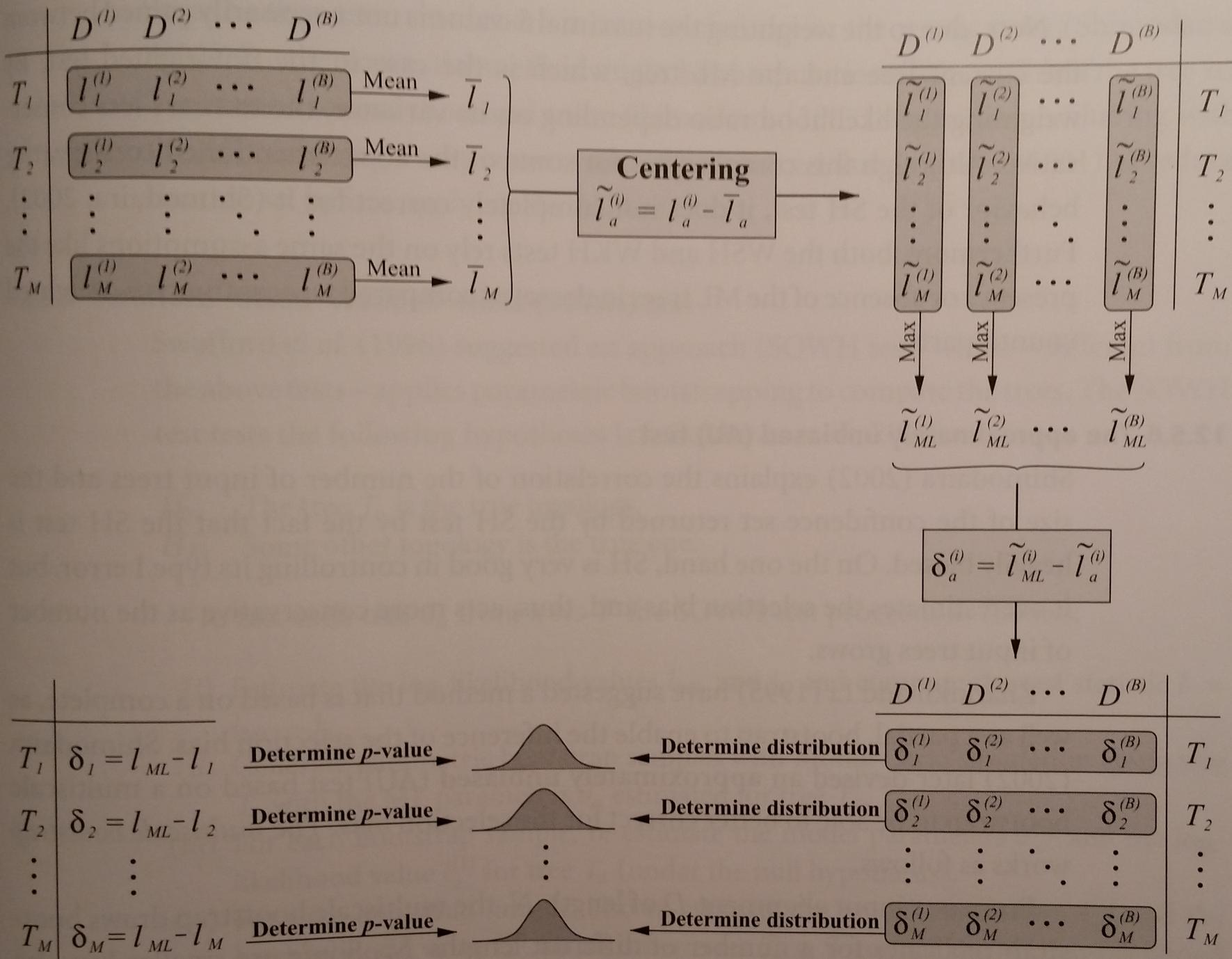
$$c_i^{(j)} = \ln L(T_i | X^{(j)}) - \bar{\ln L}(T_i | X^*)$$

- For each bootstrap replicate,  $j$ , pick the highest value from the centered distributions (this mimics the selection bias):

$$m^{(j)} = \max \left[ c_i^{(j)} \right] \text{ over all } i$$

- Then for each tree and replicate, you get a sample from the null  $\delta_i^{(j)} = m^{(j)} - c_i^{(j)}$
- $P$ -value for tree  $T_i$  is approximated by the proportions of bootstrap reps for which:

$$\delta_i^{(j)} \leq \delta(\hat{T}, T_i | X)$$



## **Another ways to assess the null distribution of the LR test statistic**

---

- Bootstrapping then centering LR, and
- Using normality assumptions.

are both clever and cute solutions.

But they do not match the null distribution under any model of sequence evolution.

## Parametric bootstrapping to generate the null distribution for the LR statistic

---

1. find the best tree and model pair that are consistent with the null,
2. Simulate many datasets under the parameters of that model,
3. Calculate  $\delta^{(j)} = 2 \left[ \ln L(\hat{T}^{(j)} | X^{(j)}) - \ln L(\hat{T}_0^{(j)} | X^{(j)}) \right]$  for each simulated dataset.
  - the  $(j)$  is just an index for the simulated dataset,
  - $\hat{T}_0^{(j)}$  is the tree under the null hypothesis for simulation replicate  $j$

## Parametric bootstrapping

---

This procedure is often referred to as SOWH test (in that form, the null tree is specified *a priori*).

Huelsenbeck et al. (1996) describes how to use the approach as a test for monophyly.

Intuitive and powerful, but not robust to model violation (Buckley, 2002).

## **KH Test**

H0: The two trees are equally supported

HA: The two trees are not equally supported

## **SH Test**

H0: All trees (including the ML tree) are equally good explanations of the data

HA: Some or all trees are not equally good explanations of the data

## **SOWH Test**

H0: Tree T from a set of trees is the true topology

HA: Some other topology is the true one

Let's run tree-puzzle to perform the KH and SH tests...

```
$ tree-puzzle PHYLIP.Alignment PHYLIP.tre
```