Testing phylogenetic hypotheses

Woods Hole Workshop on Molecular Evolution, 2015

Mark T. Holder
University of Kansas

Thanks to Paul Lewis, Joe Felsenstein, and Peter Beerli for slides.

Reasons phylogenetic inference might be wrong

1. *Systematic error* – Our inference method might not be sophisticated enough
2. *Random error* – We might not have enough data – we are misled by sampling error.

(or it could be some combination of these).

Focus of this lecture: **How confident can we be in the trees/splits inferred by ML?**

1. Bootstrapping
2. Putting P-values on trees:
   - KH Test, SH Test
   - parametric bootstrapping,
   - aLRT, aBayes,
   - 1 - BP,
   - AU and Efron et al. (1996) correction
   - aBP
3. More *caveats*

Bayesian vs frequentist statistics

The Bayesian approach is quite different – it even entails a different definition of probability!
Sometimes the difference is portrayed as an intractable philosophical battle.
I think that of it as a different way to make arguments.

"Politicians use statistics in the same way that a drunk uses lamp-posts — for support rather than illumination."
Andrew Lang

- We should be using statistics for our own illuminaton, but we should also think of statistics as a communication tool.
- Don't perform *only* the analyses that would convince *you*, make sure that your claims are convincing to a wide range of viewpoints.

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument).

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument).
2. If we used this procedure a lot, we'd bracket the true value 95% of the time (the confidence interval)

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument).
2. If we used this procedure a lot, we'd bracket the true value 95% of the time (the confidence interval).
3. If $\mathbb{P}(\theta)$ describes our beliefs about plausible models before seeing the data, D, after we see the data our beliefs should be $\mathbb{P}(\theta \mid D)$. The Bayesian posterior probability.

The (relatively few) types of arguments we make with statistics

1. These data would be unlikely if the null were true (the p-value argument),
2. If we used this procedure a lot, we'd bracket the true value 95% of the time (the confidence interval)
3. If $\mathbb{P}(\theta)$ describes our beliefs about plausible models before seeing the data, D, after we see the data our beliefs should be $\mathbb{P}(\theta \mid D)$. The Bayesian posterior probability.
4. If we use this threshold for evidence, we would expect 5% of our positives to be false positives (the false discovery rate)

The p-value argument

These data would be unlikely if the null were true.
Pros:

1. Very conservative. Gives the null the benefit of the doubt by focusing on the "least favorable" conditions.
2. If you don't reject the null, then your conclusions resistant to the addition of more models.

Cons:

1. Can be surprisingly difficult to correctly calculate p
2. Does not give you inference to the best model.
3. Failure to reject null often caused by lack of data.

The confidence interval argument

Pros:

1. Somewhat intuitive.
2. Identifies a plausible set of answers.
3. Not dependent on prior knowledge of parameters.

Cons:

1. Can be surprisingly difficult to correctly calculate
2. Ignores prior information
3. Does not fully condition on the data you observed.

The Bayesian posterior probability argument. $\mathbb{P}(\theta|X)$.

Pros:

1. Very intuitive statement of the best range of parameters
2. The only coherent statement of knowledge that uses all of the information in the data

Cons:

1. Unconvincing to people with different priors
2. Its unclear what would happen if you consider another model
3. Relies on MCMC (which is cool, but dangerous)

The False discovery rate argument.

Pros:

1. Nice mixture of frequentist and emprical Bayesian behavior
2. Check out Nicolas Lartillot's blog `http://bayesiancook.blogspot.com/` for convincing arguments that evolutionary genomics has an empirical Bayesian future.

Cons:

1. You're not going to see a lot of software in phylogenetics that spits out FDR values.

For each analysis/test that we talk about in the course:

1. make sure that you understand the "signal":
   - sketch out cases of the tree or data that would look uninteresting,
   - sketch what interesting data would look like
2. What confounding factors could lead to similiar signal?

   - *e.g.* analyses of diversification times and rates depend crucially on branch lengths.
   - models that are too simplistic distort branch lengths (mainly they underestimate deep and long branches)

Some resources related to this talk

A Zotero Group of papers related to topology testing on trees.

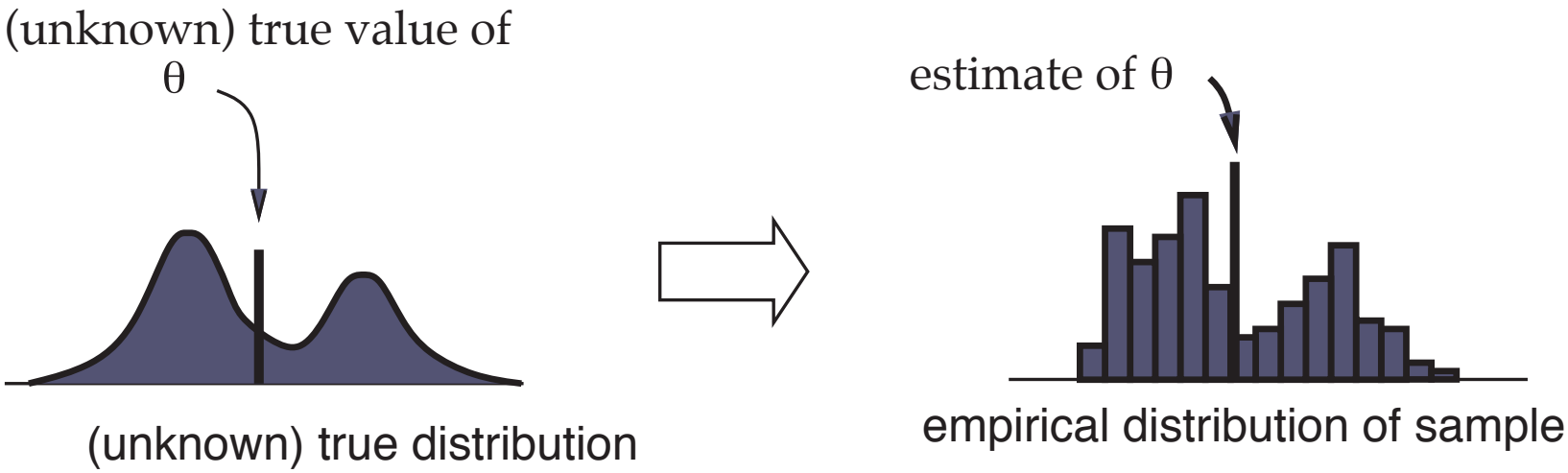
A http://phylo.bio.ku.edu/woodshole/index.html has the beginnings of an annotated bibliography and some other notes.

The source for all the documents for my talk are at:

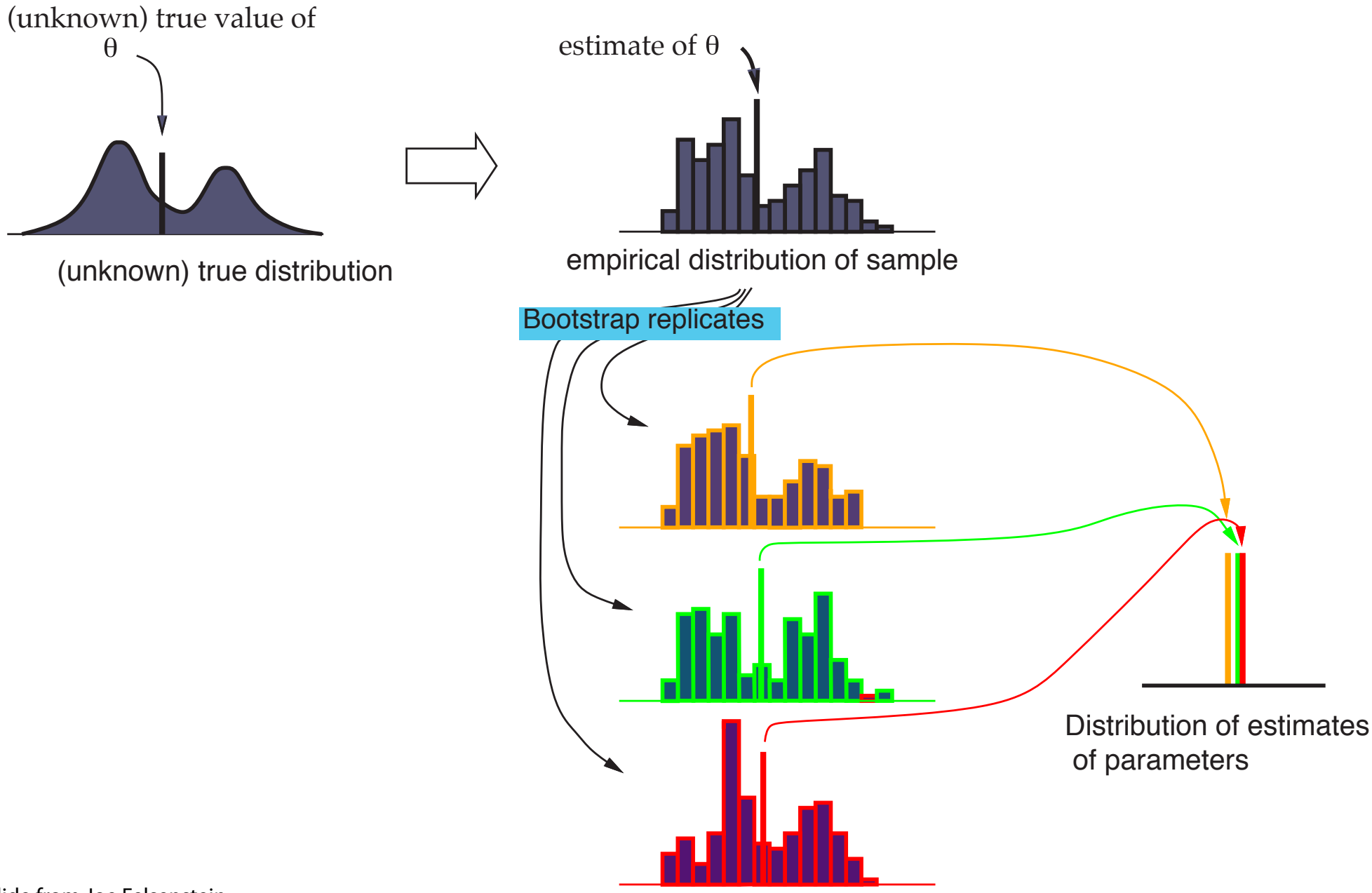
https://github.com/mtholder/TreeTopoTestingTalks

https://github.com/mtholder/treeTestingDemo

The bootstrap

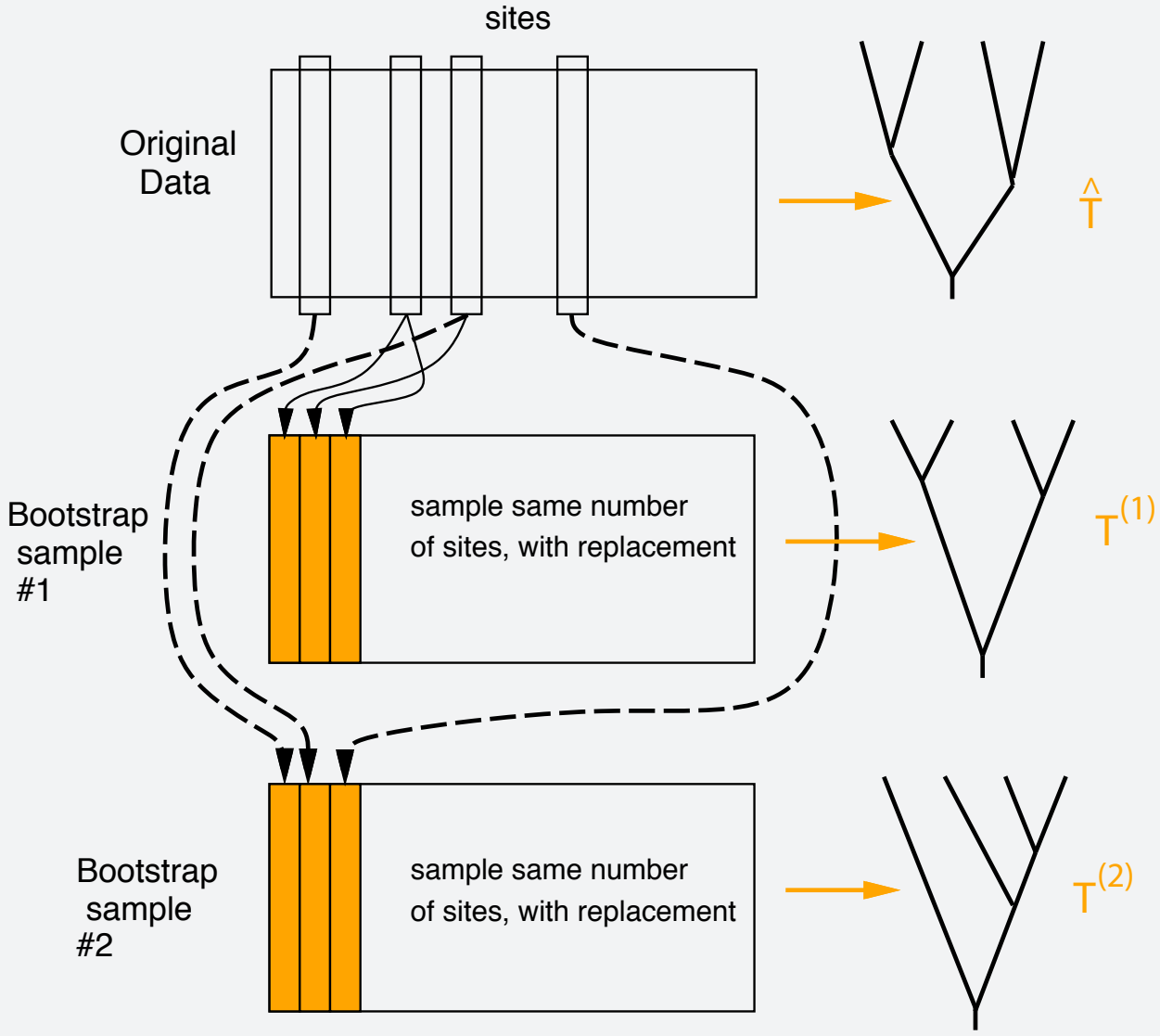


The bootstrap



Slide from Joe Felsenstein

The bootstrap for phylogenies

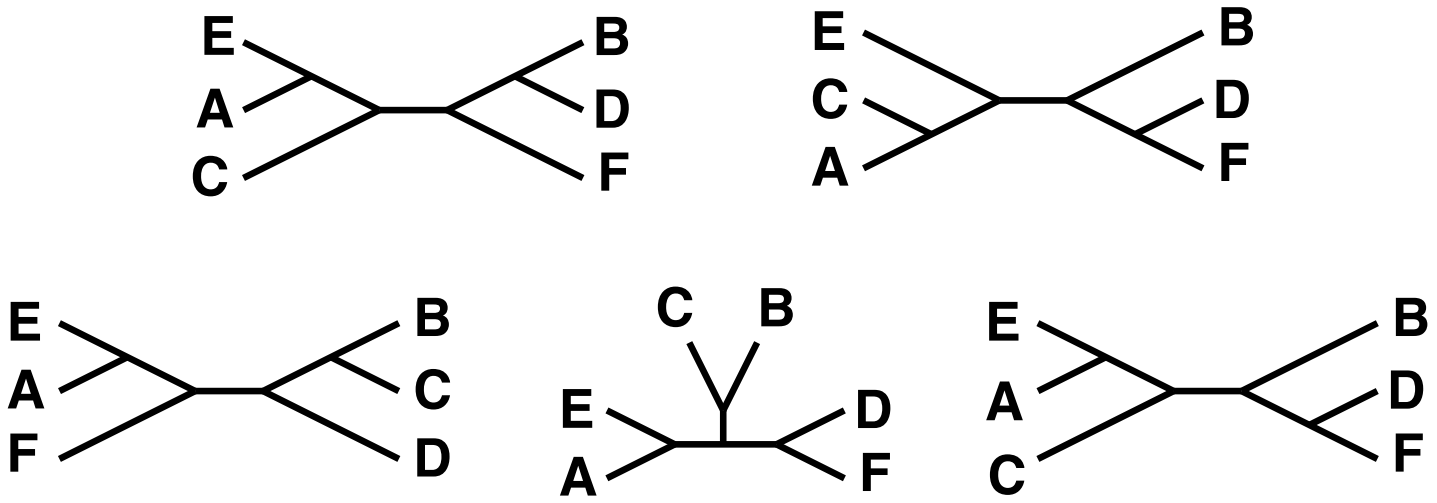


Slide from Joe Felsenstein

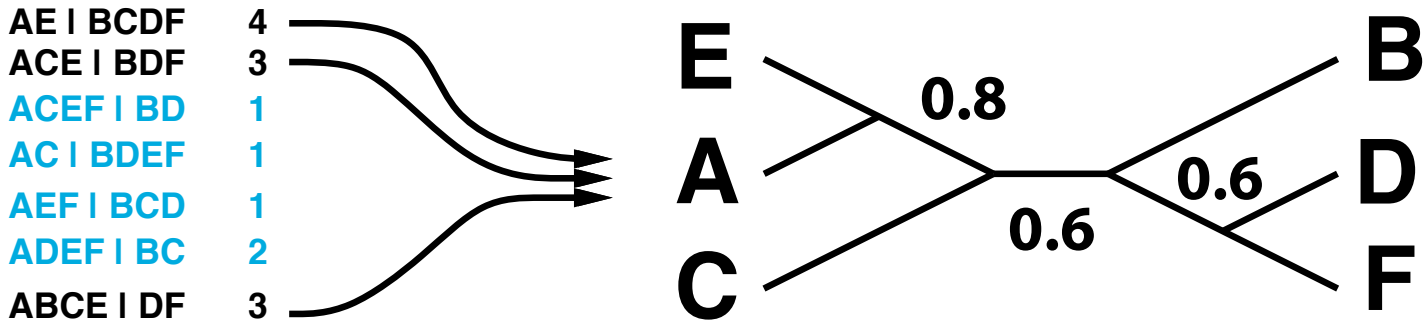
(and so on)

The majority-rule consensus tree

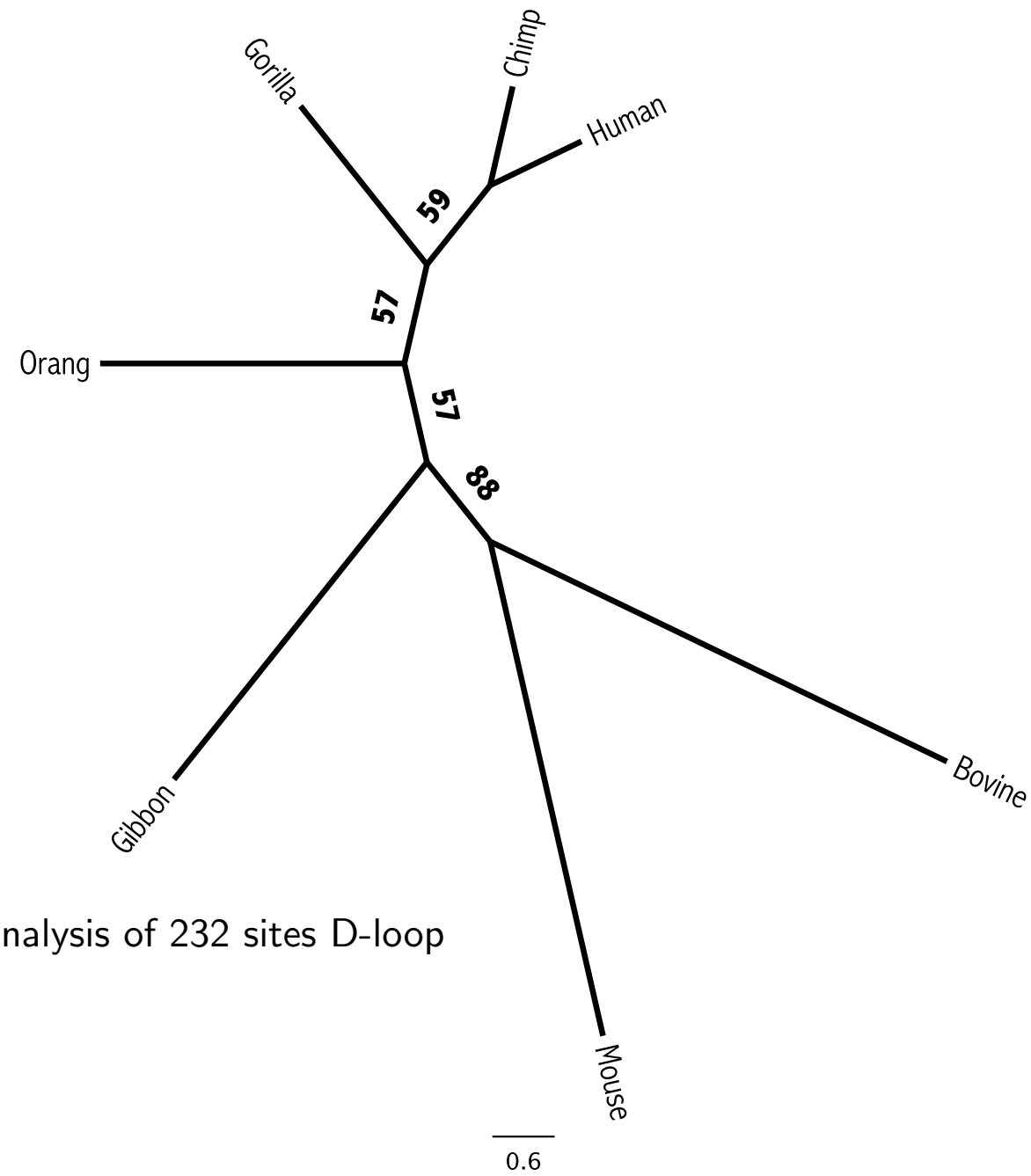
Trees:



How many times each partition of species is found:



Slide from Joe Felsenstein



From Hasegawa's analysis of 232 sites D-loop

http://phylo.bio.ku.edu/mephytis/boot-sample.html

http://phylo.bio.ku.edu/mephytis/parsimony.html

http://phylo.bio.ku.edu/mephytis/bootstrap.html

Bootstrapping for branch support

- Typically a few hundred bootstrap, pseudoreplicate datasets are produced.

- Less thorough searching is faster, but will usually artificially lower bootstrap proportions (BP). However, Anisimova et al. (2011) report that RAxML's rapid bootstrap algorithm may inflate BP.

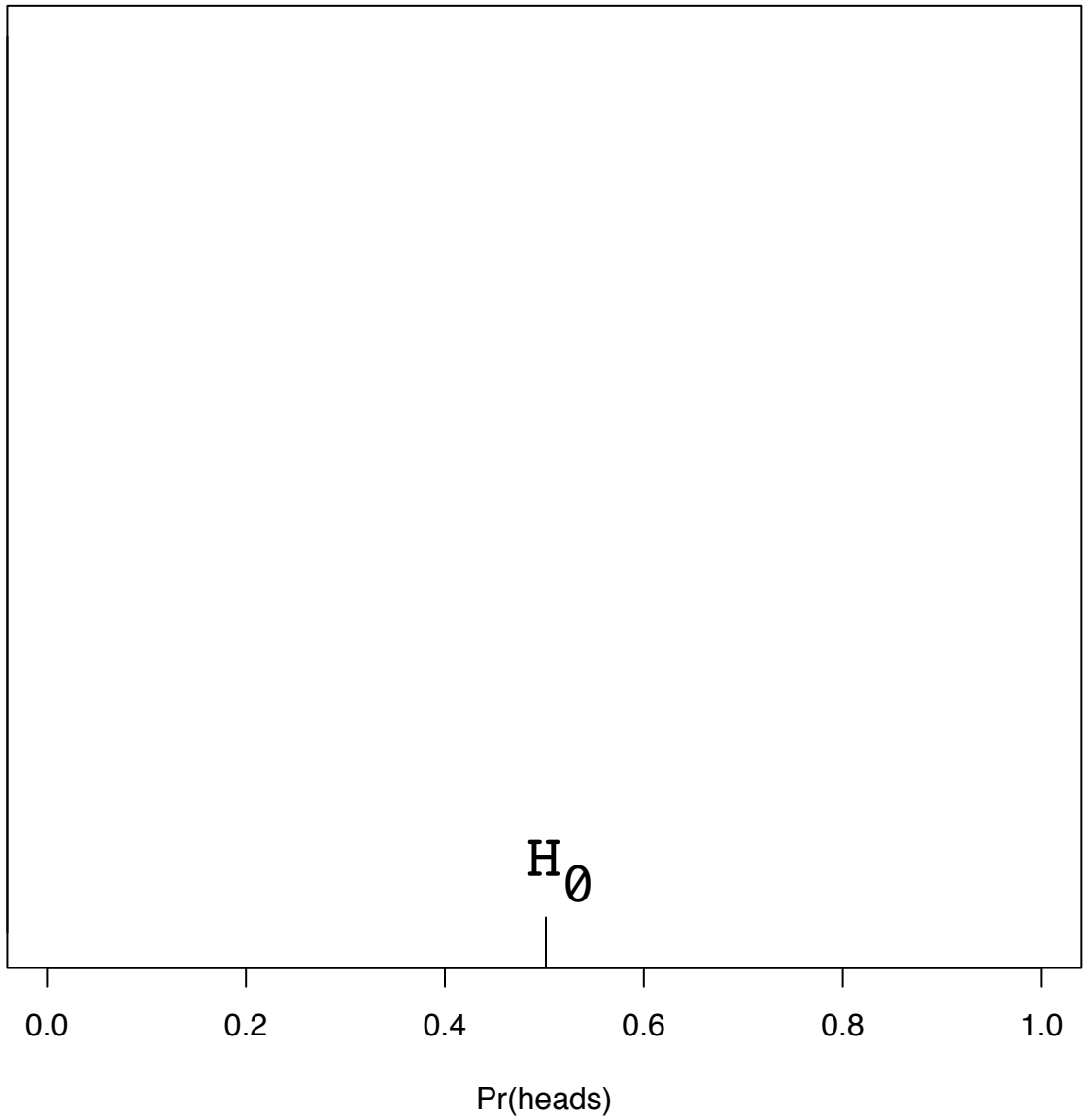
- "Rogue" taxa can lower support for many splits – you do not have to use the majority-rule consensus tree to summarize bootstrap confidence statements.

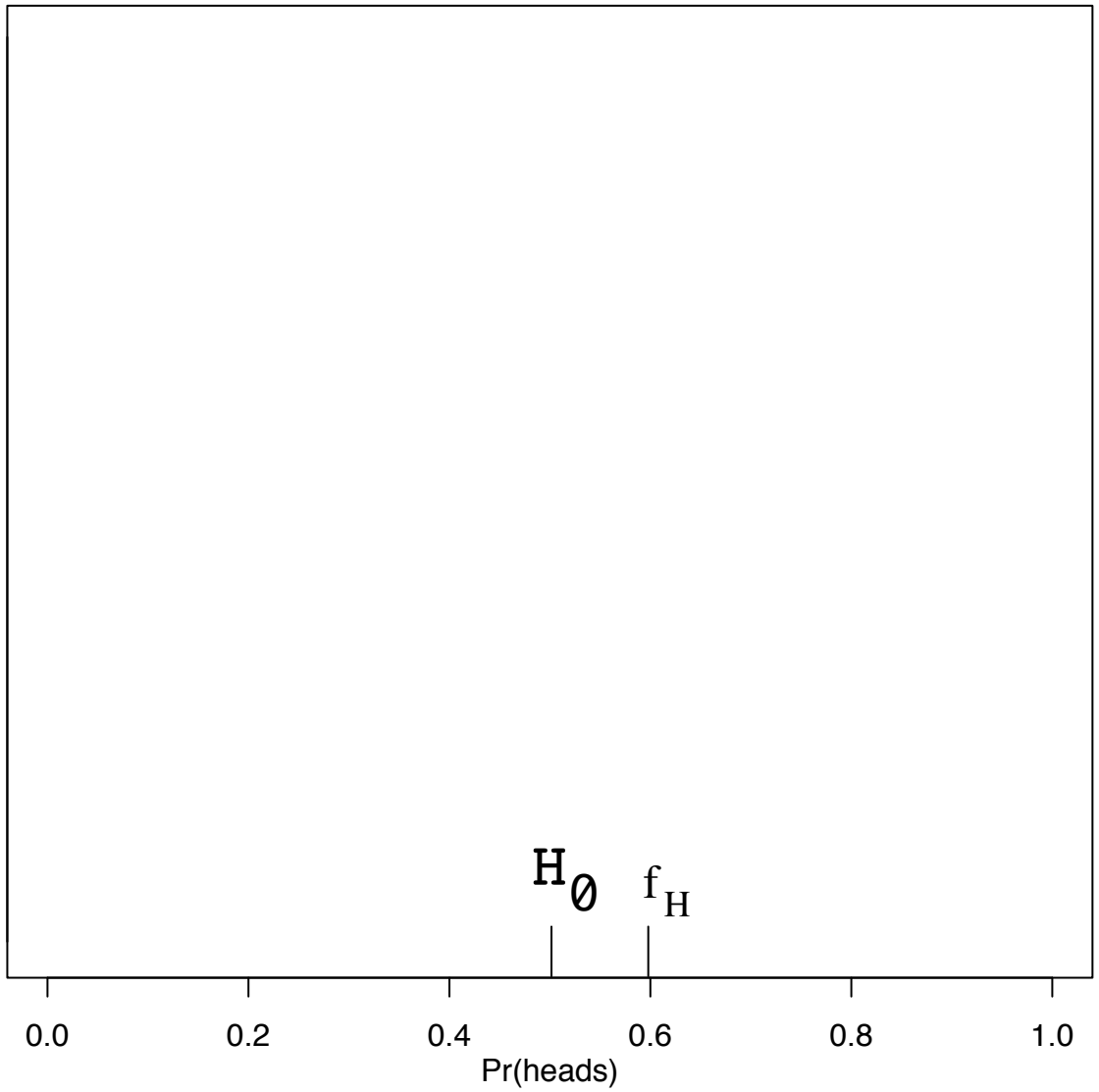
Frequentist hypothesis testing: coin flipping example

$N = 100$ and $h = 60$
Can we reject the fair coin hypothesis? $H_0 : \Pr(\text{heads}) = 0.5$

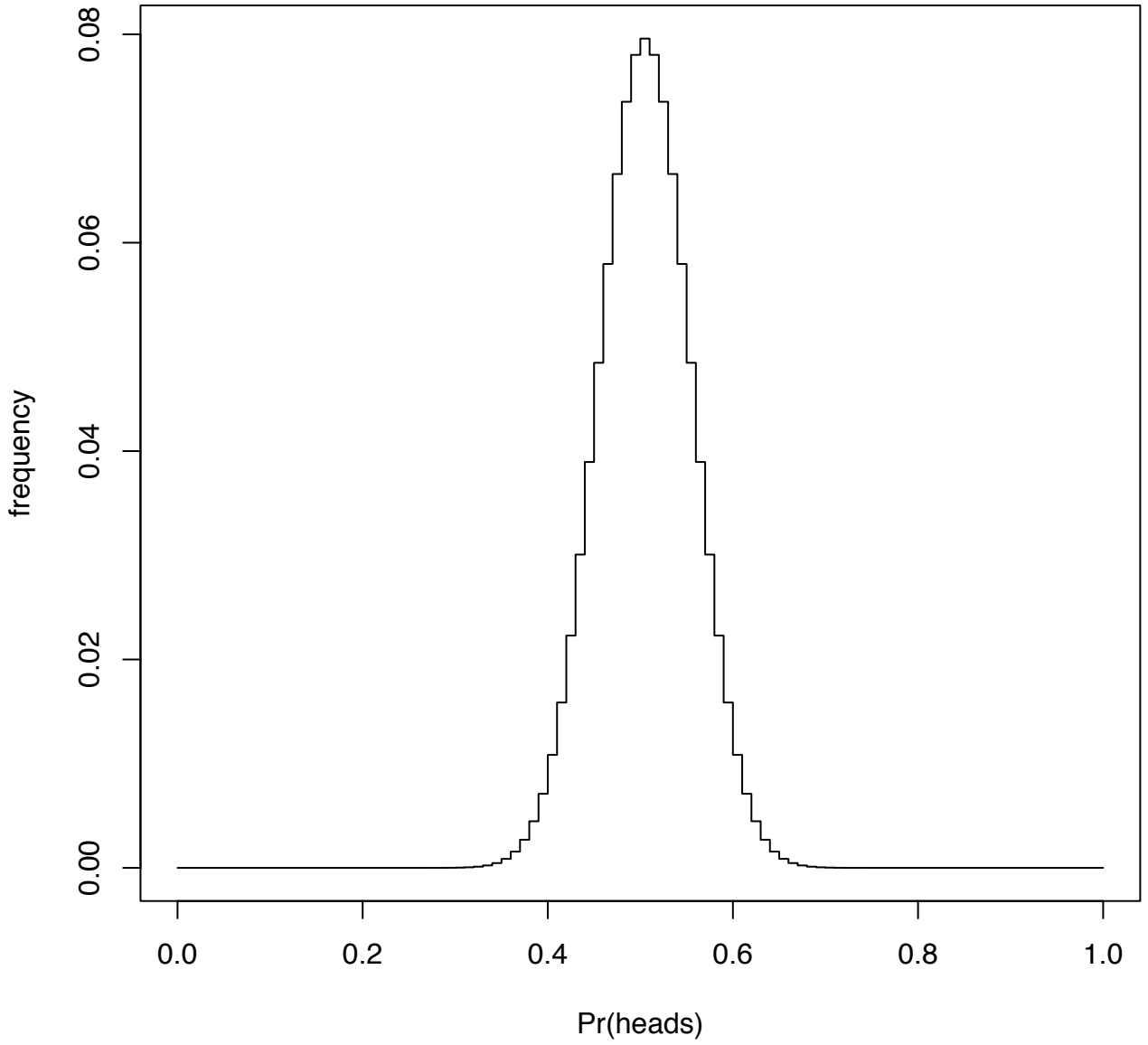
The "recipe" is:

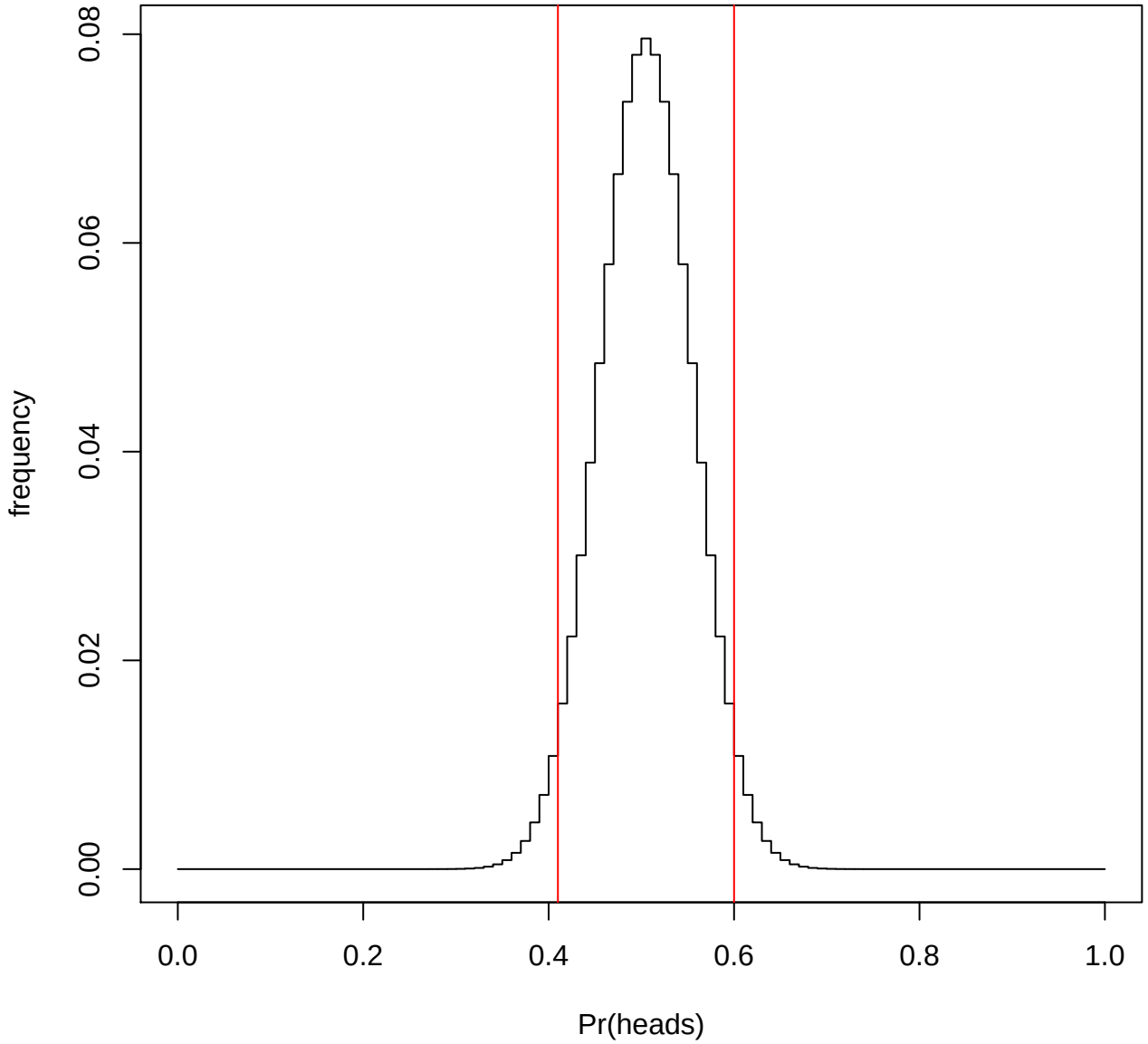
1. Formulate null (H_0) and alternative (H_A) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic: $f_H =$ fraction of heads in sample. $f_H = 0.6$
4. Characterize the null distribution of the test statistic
5. Calculate the P-value: The probability of 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.


H_0
0.0
0.2
0.4
0.6
0.8
1.0
Pr(heads)


H0
fH
0.0
0.2
0.4
0.6
0.8
1.0
Pr(heads)

Null distribution





P-value ≈ 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.
- We will see later (during "cartoon time"), that we *can* visualize them both in a parameter space that describes the frequency of different data patterns.

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:

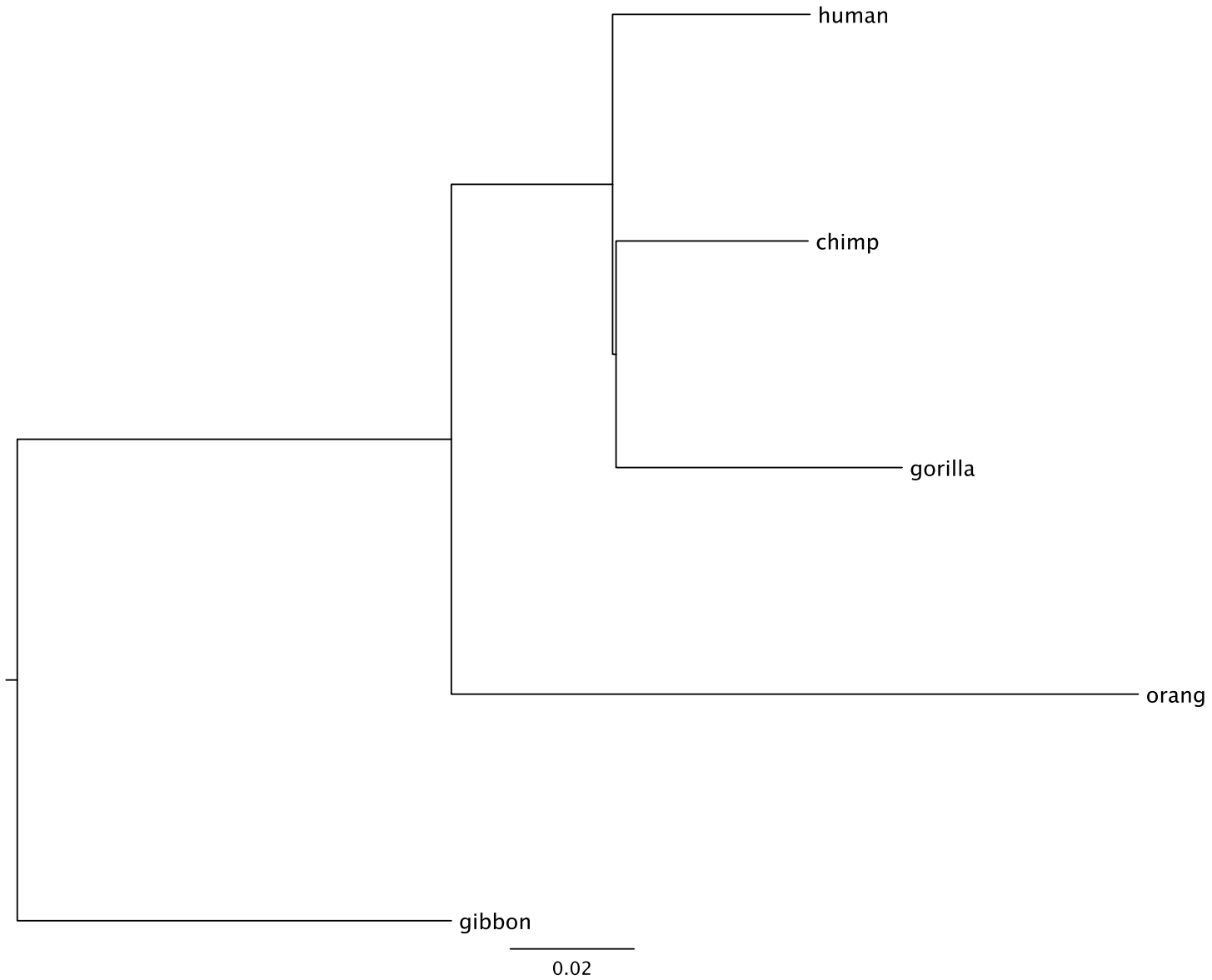
$$\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$$

Expectation under null:

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

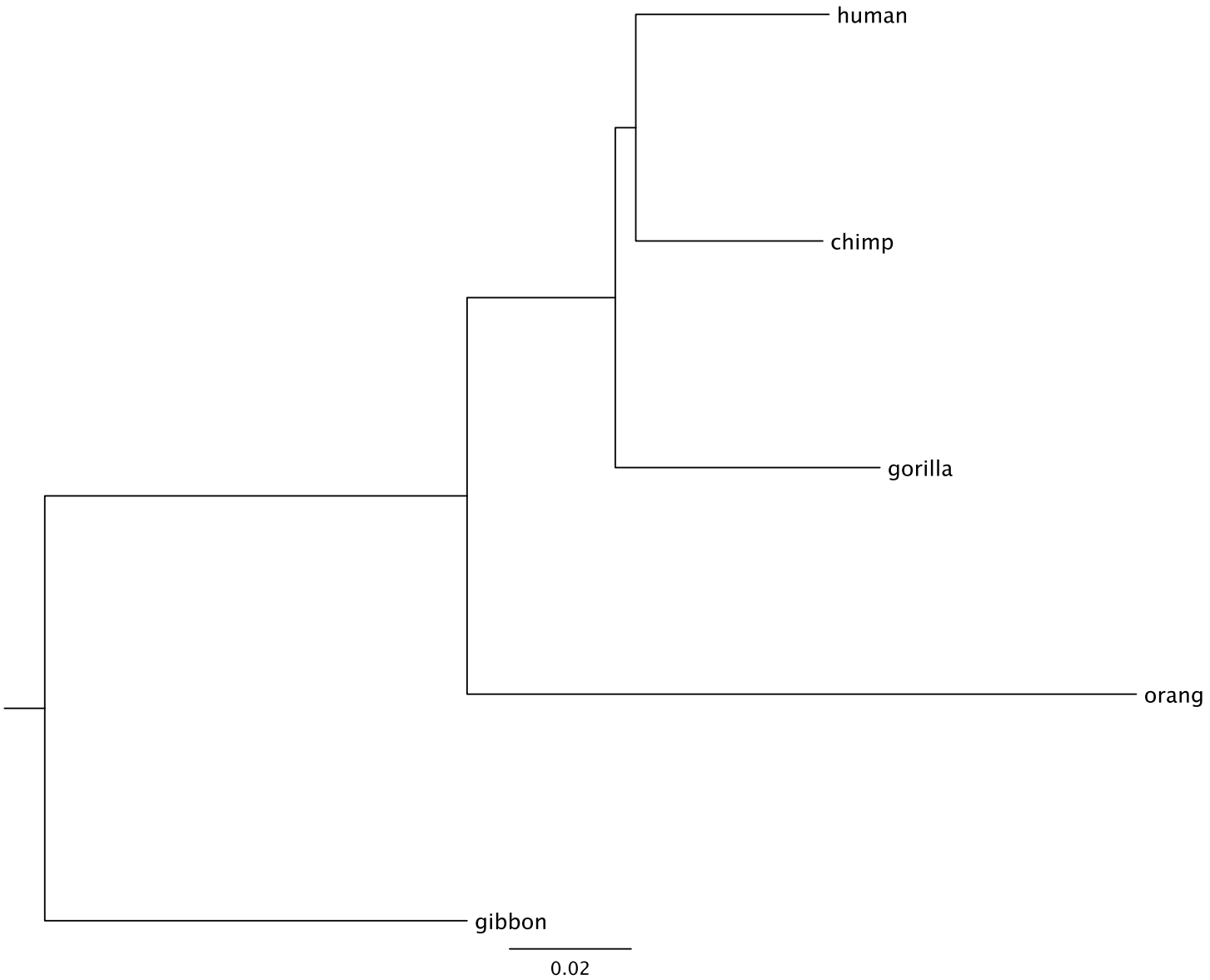
Using 3000 sites of mtDNA sequence for 5 primates

T_1 is ((chimp, gorilla), human)



Using 3000 sites of mtDNA sequence for 5 primates

T_2 is ((chimp, human), gorilla)

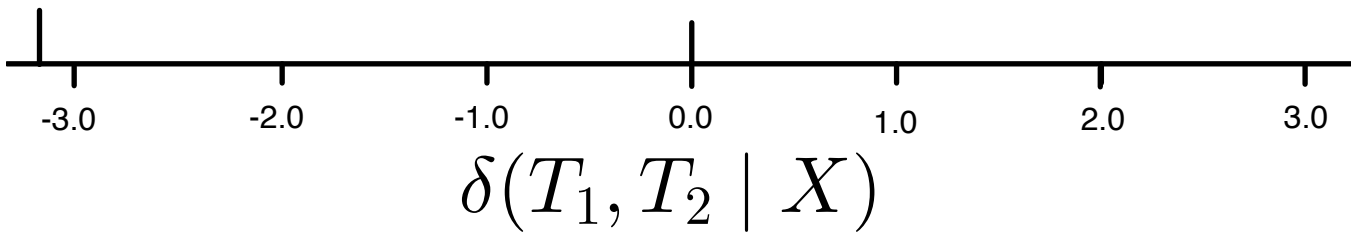


Using 3000 sites of mtDNA sequence for 5 primates

T_1 is ((chimp, gorilla), human) $\quad \ln L(T_1 \mid X) = -7363.296$

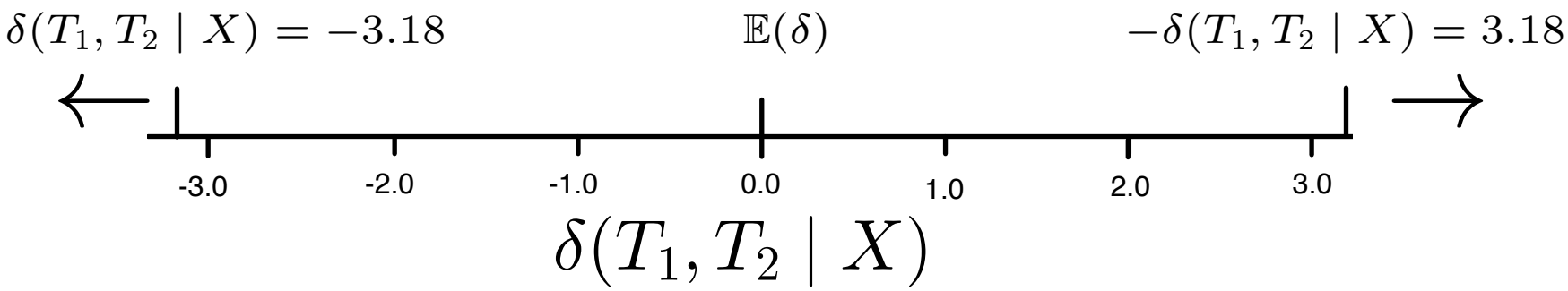
T_2 is ((chimp, human), gorilla) $\quad \ln L(T_2 \mid X) = -7361.707$

$\delta(T_1, T_2 \mid X) = -3.18$ $\quad \mathbb{E}(\delta)$



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

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



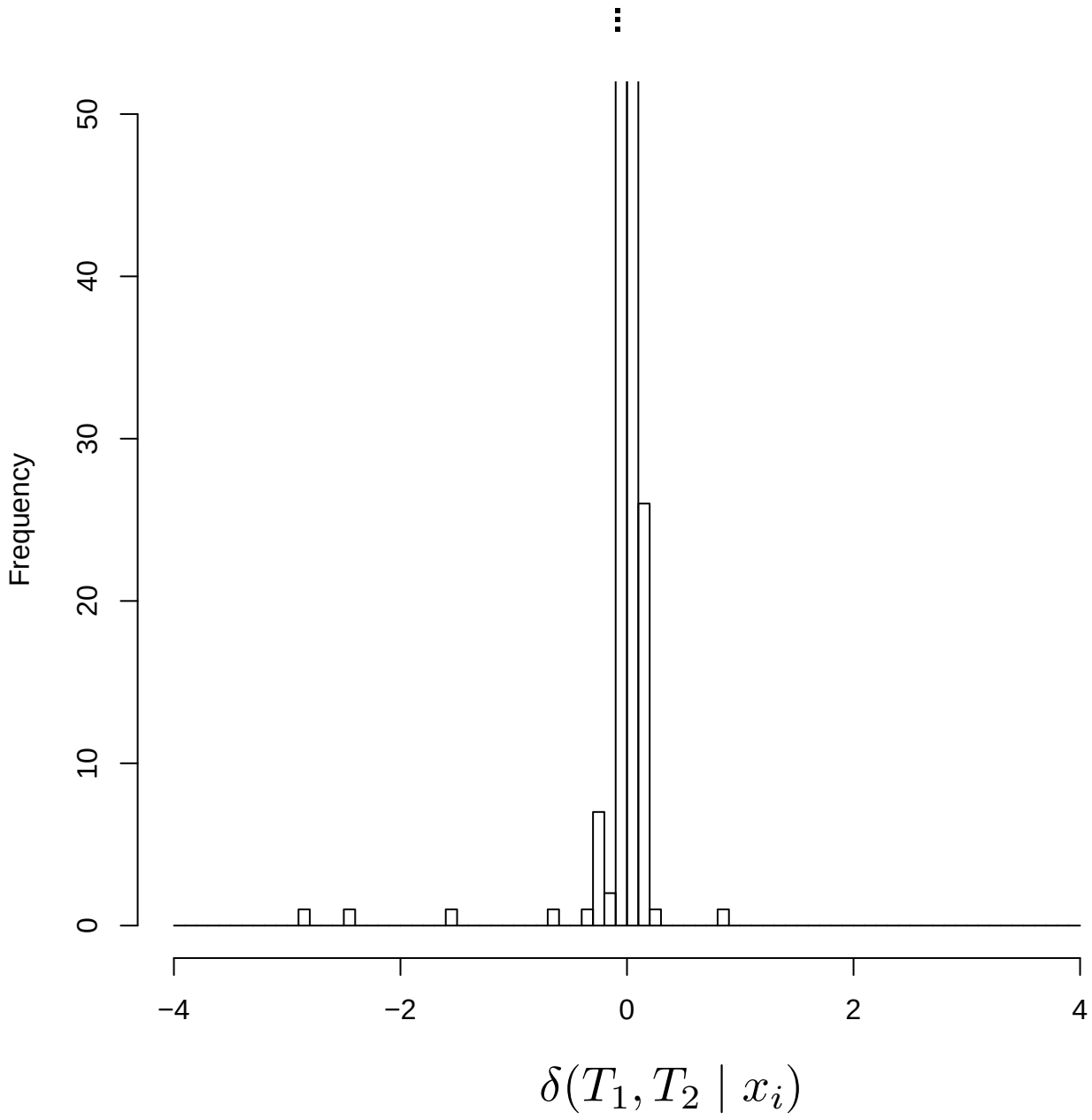
KH Test

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

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

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

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



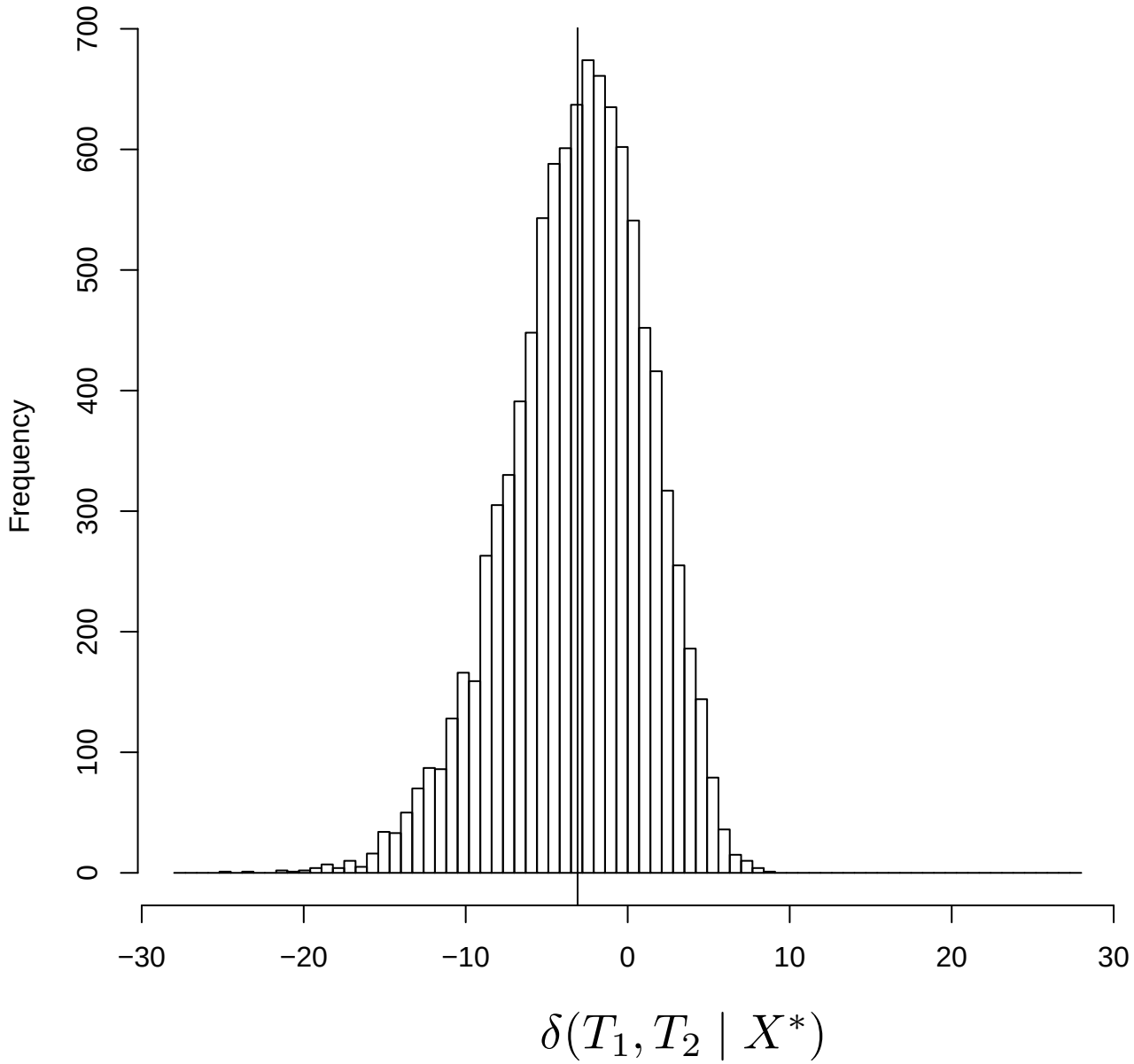
KH Test - the variance of $\delta(T_1, T_2 \mid X)$

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

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

[1] Susko (2014) recently showed that this is flawed and too conservative.

δ 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.

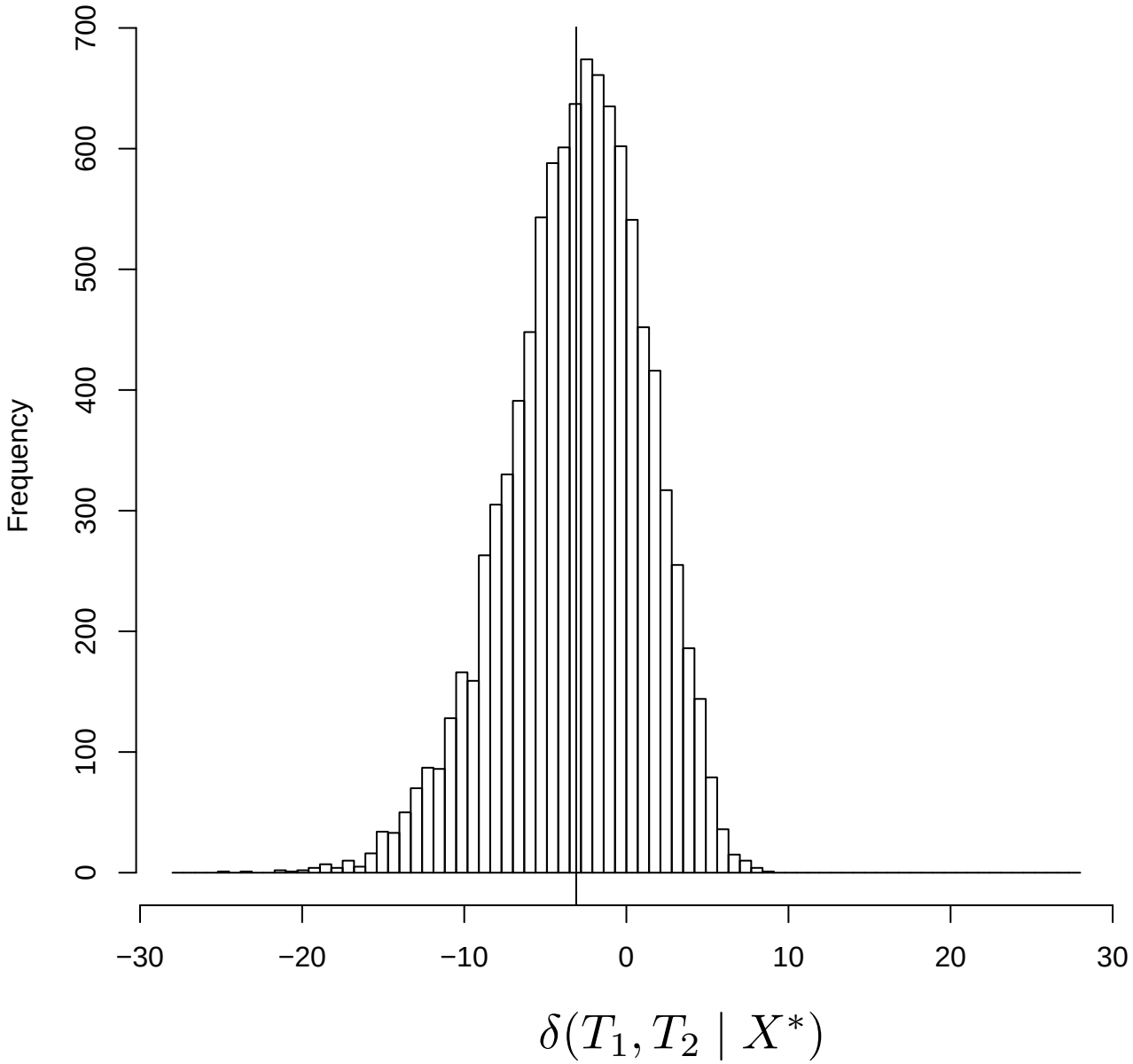
So, we can simply resample the site $\ln L$ values and sum them (rather than reoptimizing parameters).

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 δ test statistic?



KH Test - 'centering'

H_0 gives us the expected value:

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

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.

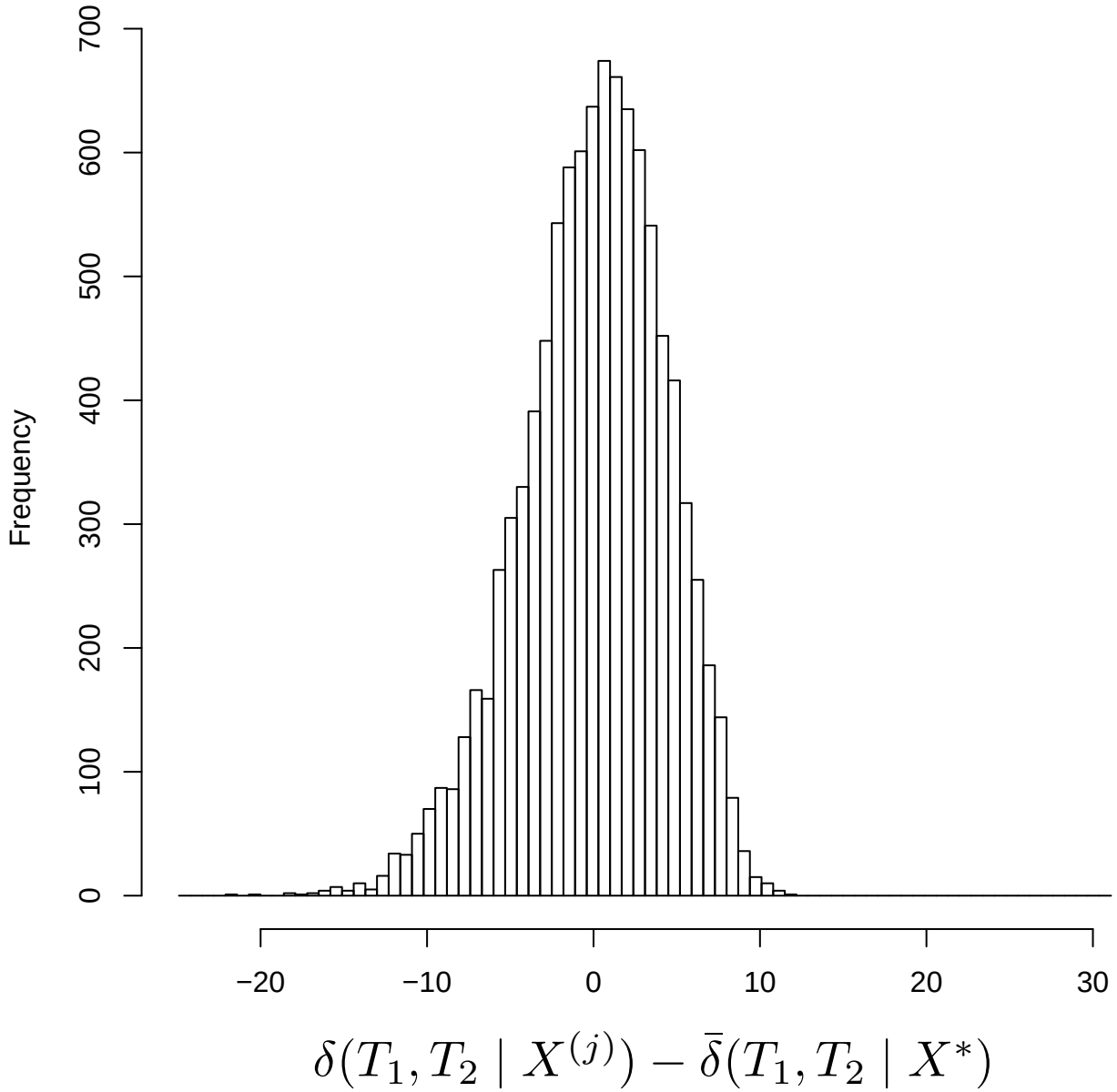
For each of the j bootstrap replicates, we treat

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

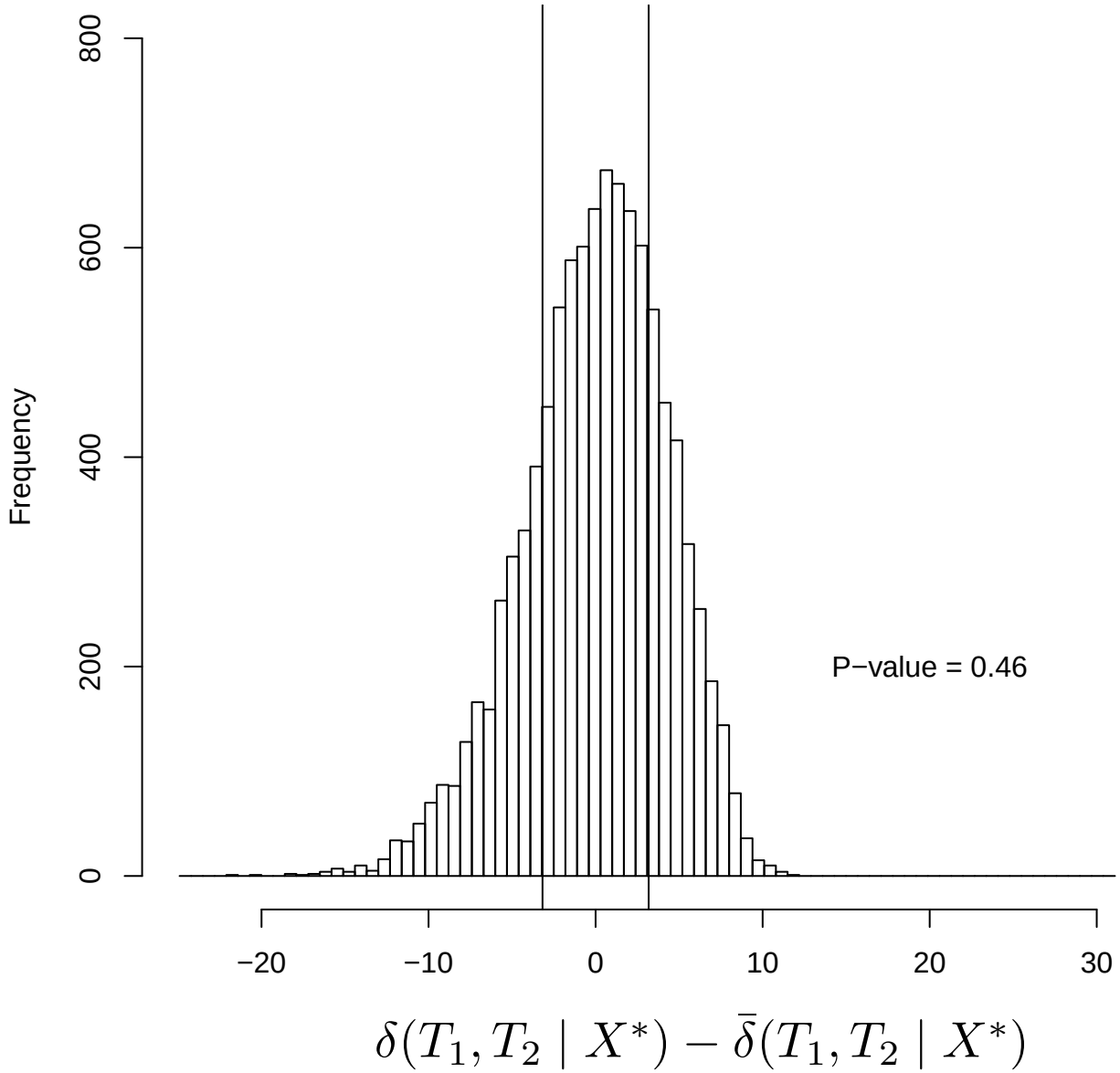
as draws from the null distribution.

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

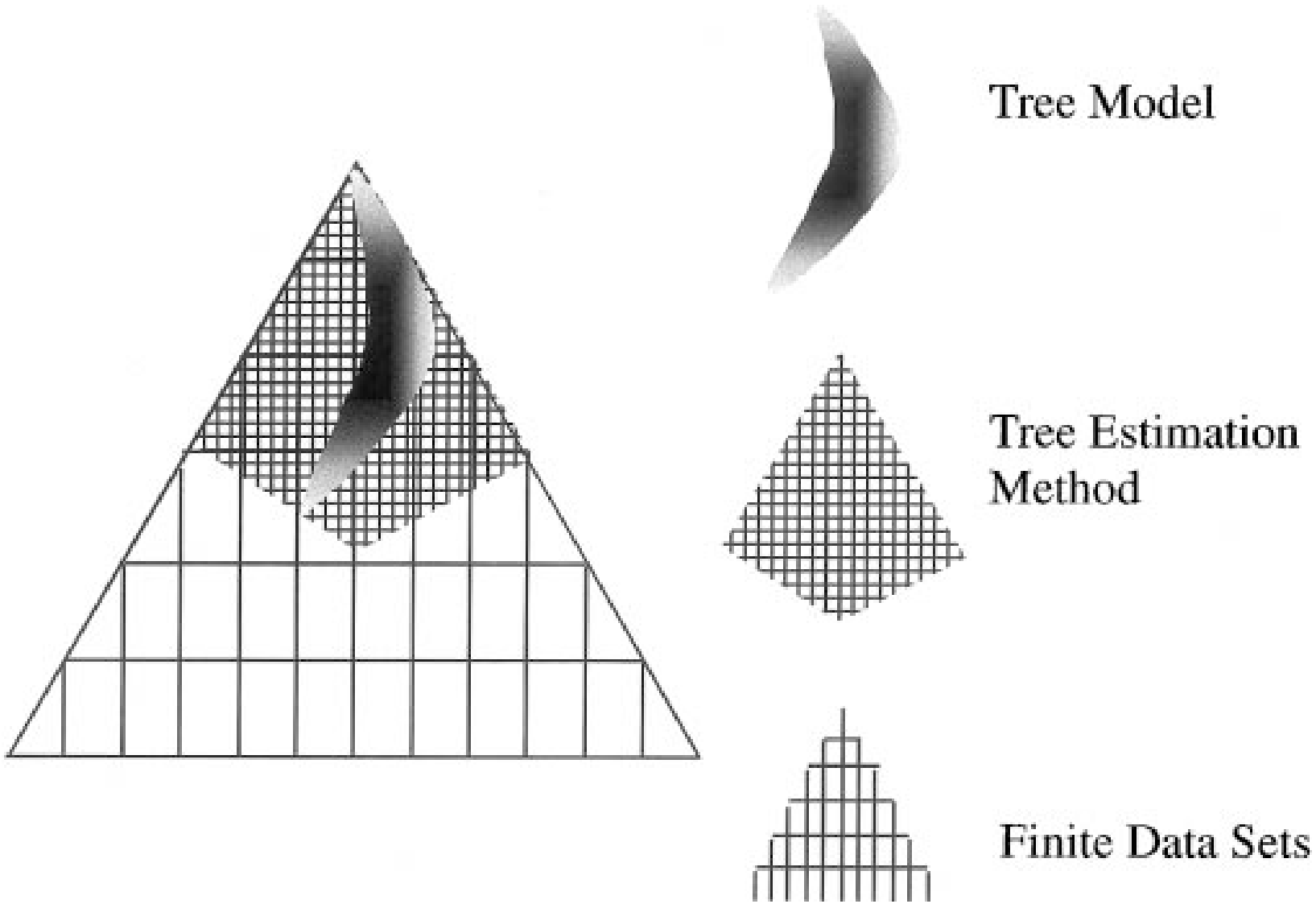
for many (RELL) bootstrapped replicates of the data



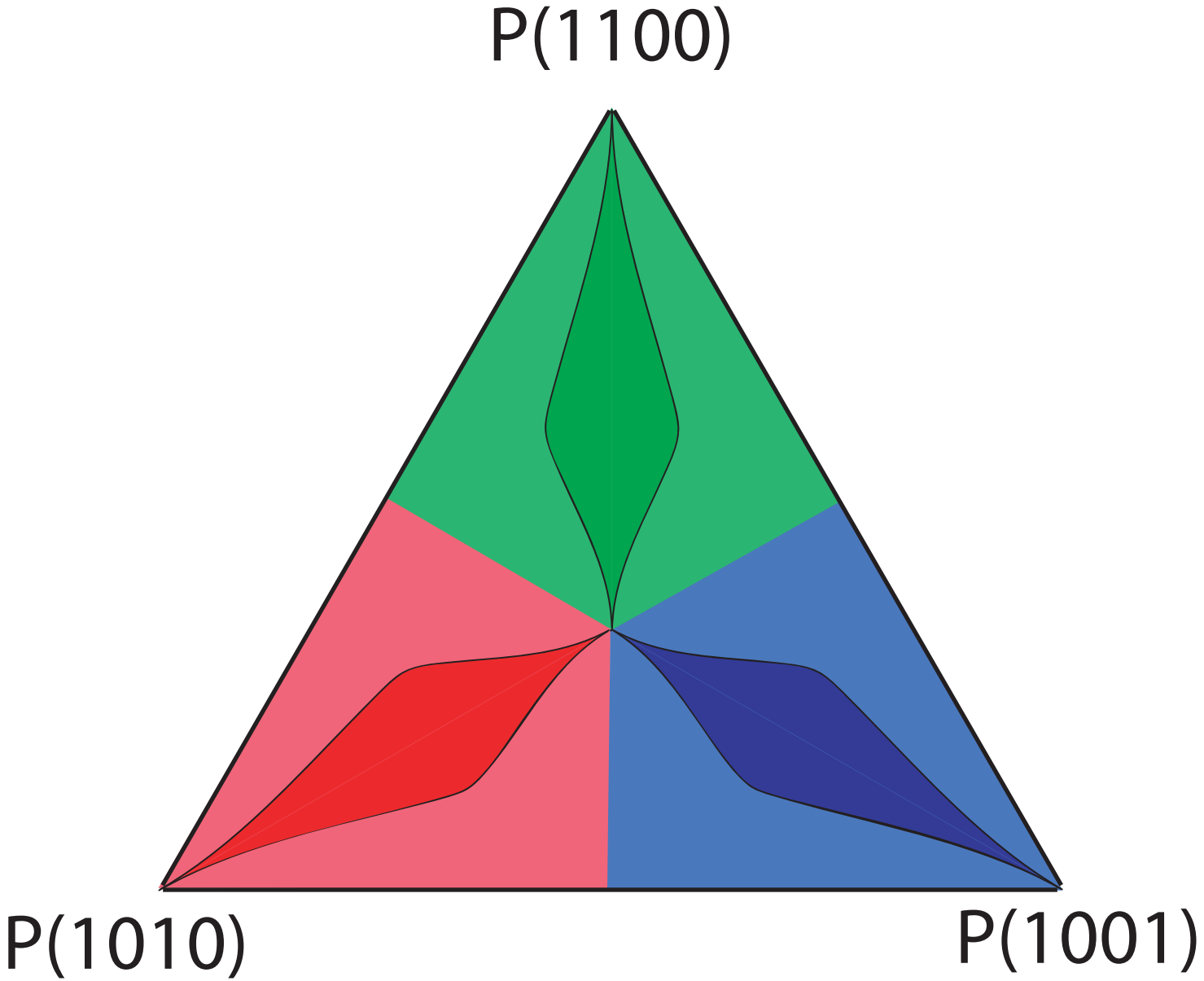
Approximate null distribution with tails (absolute value ≥ 3.18) shown



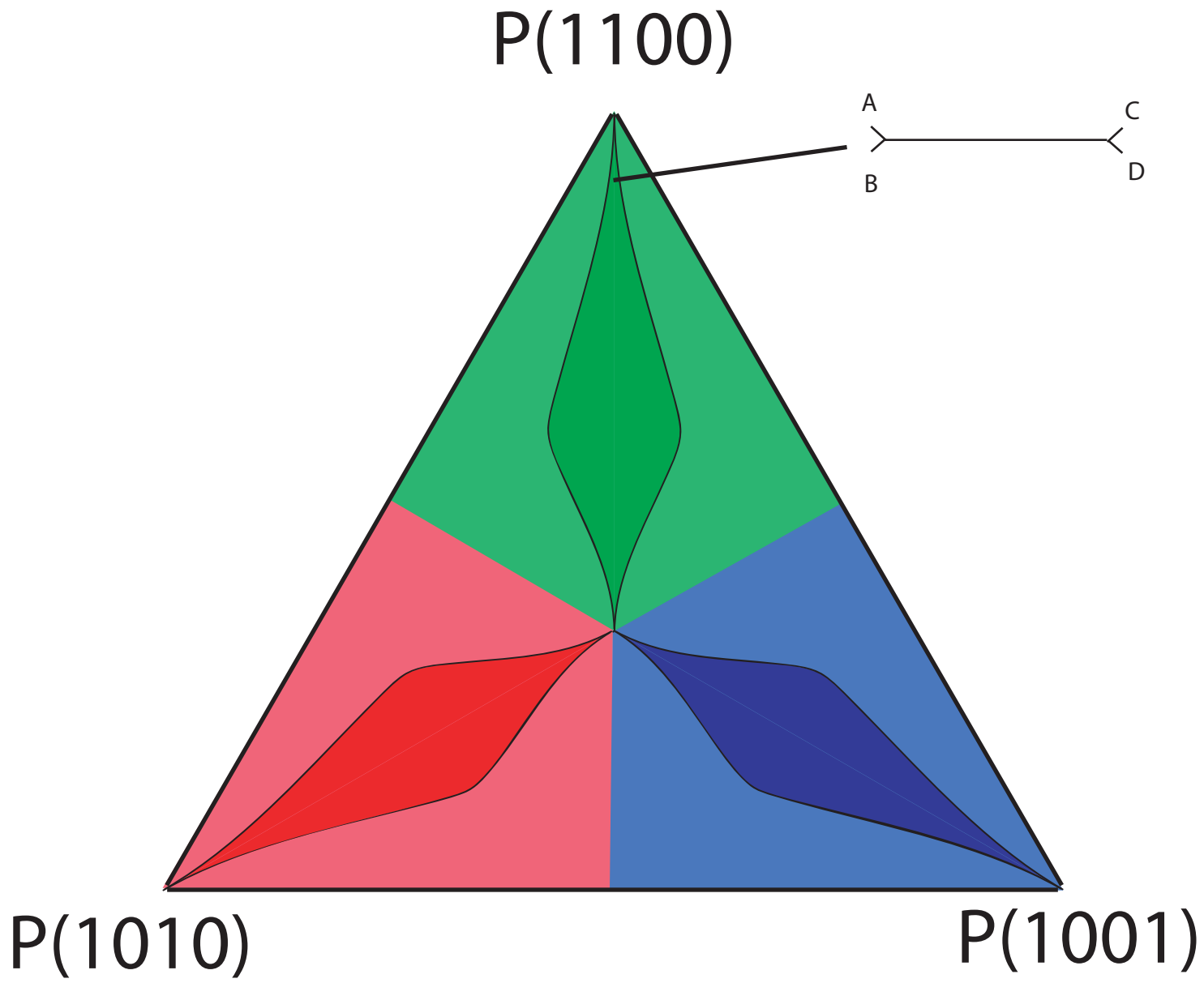
Cartoon time courtesy of the Kim (2000) view of tree space



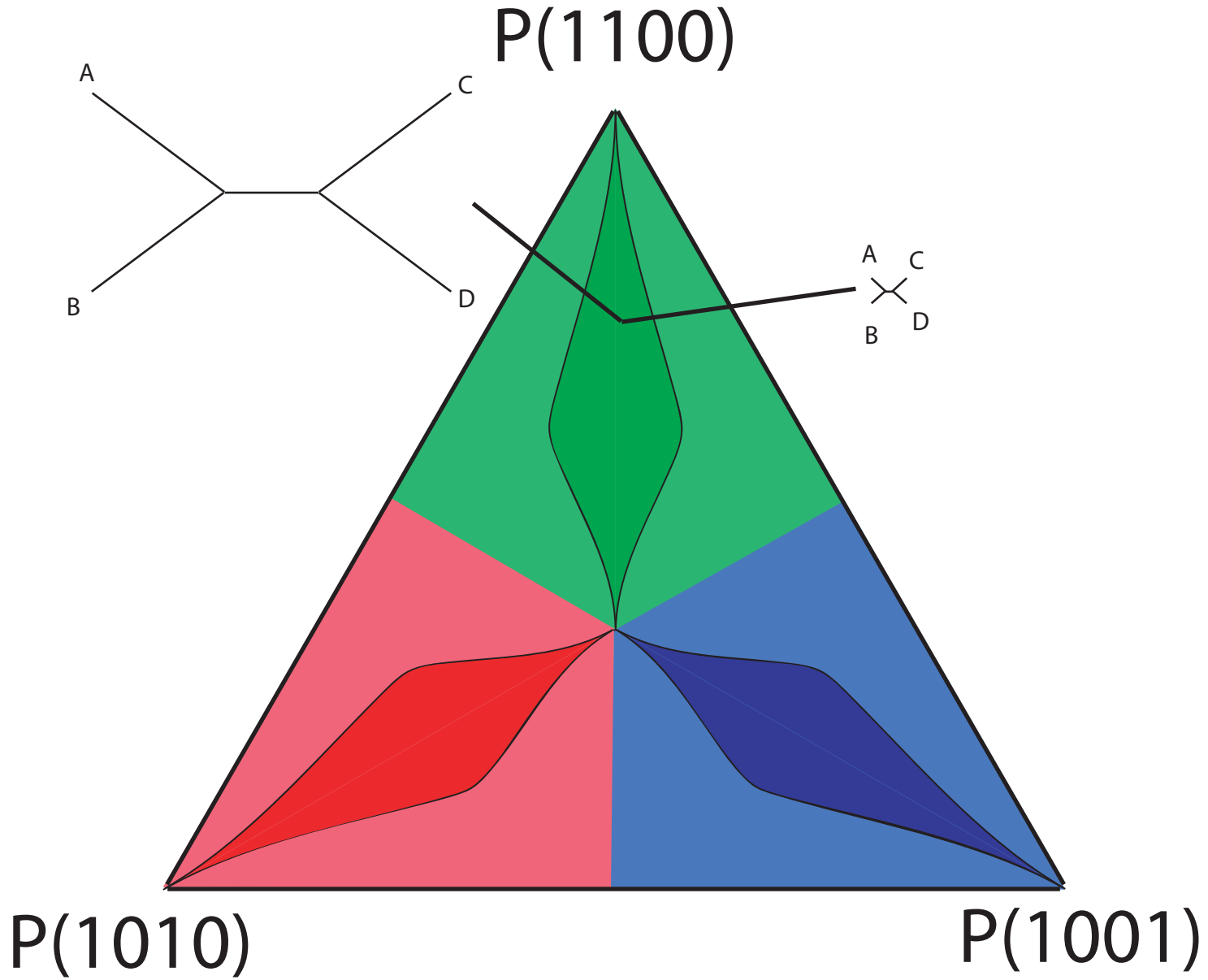
Parsimony-informative Pattern Frequency Space



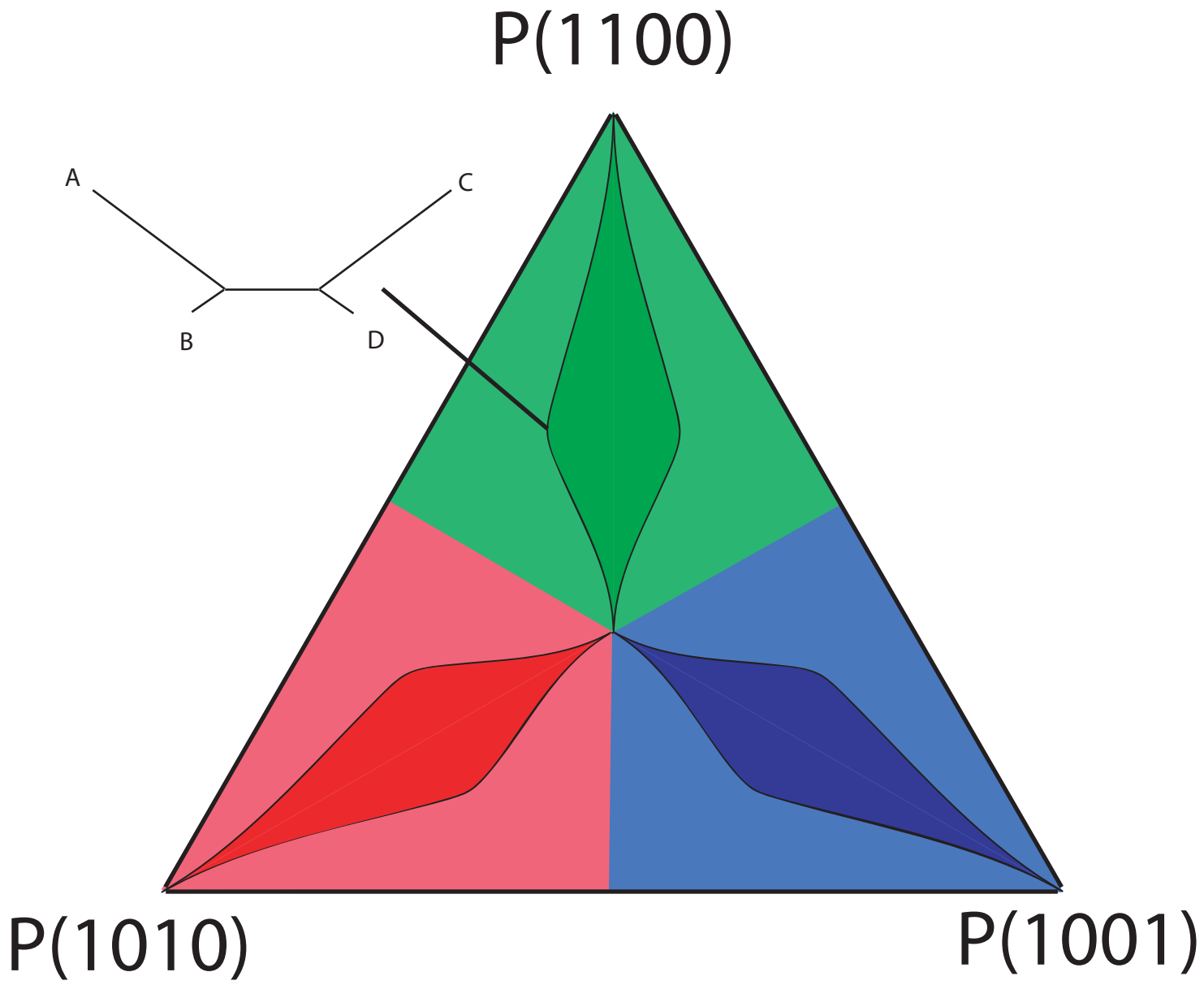
Parsimony-informative Pattern Frequency Space



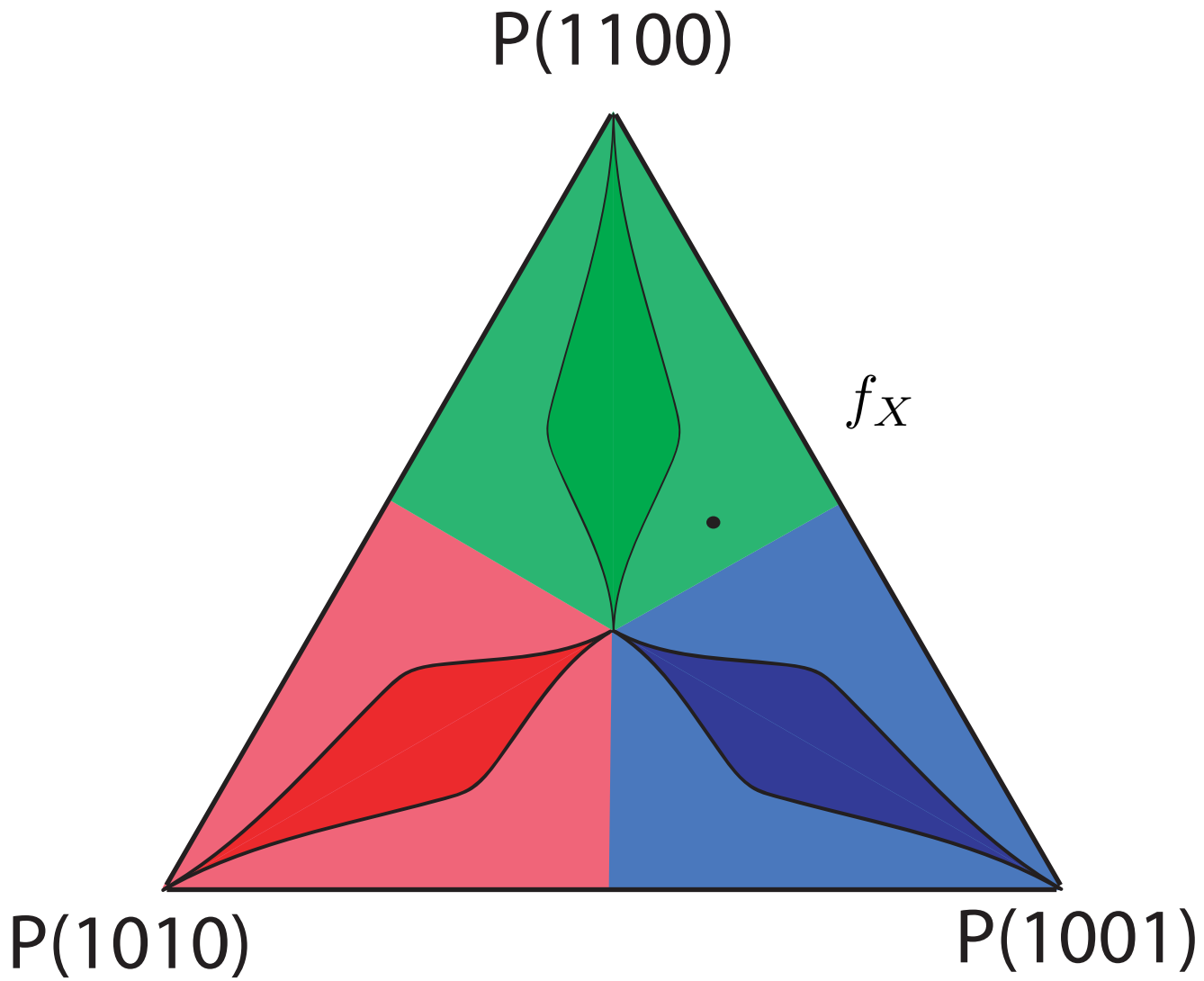
Parsimony-informative Pattern Frequency Space



Parsimony-informative Pattern Frequency Space

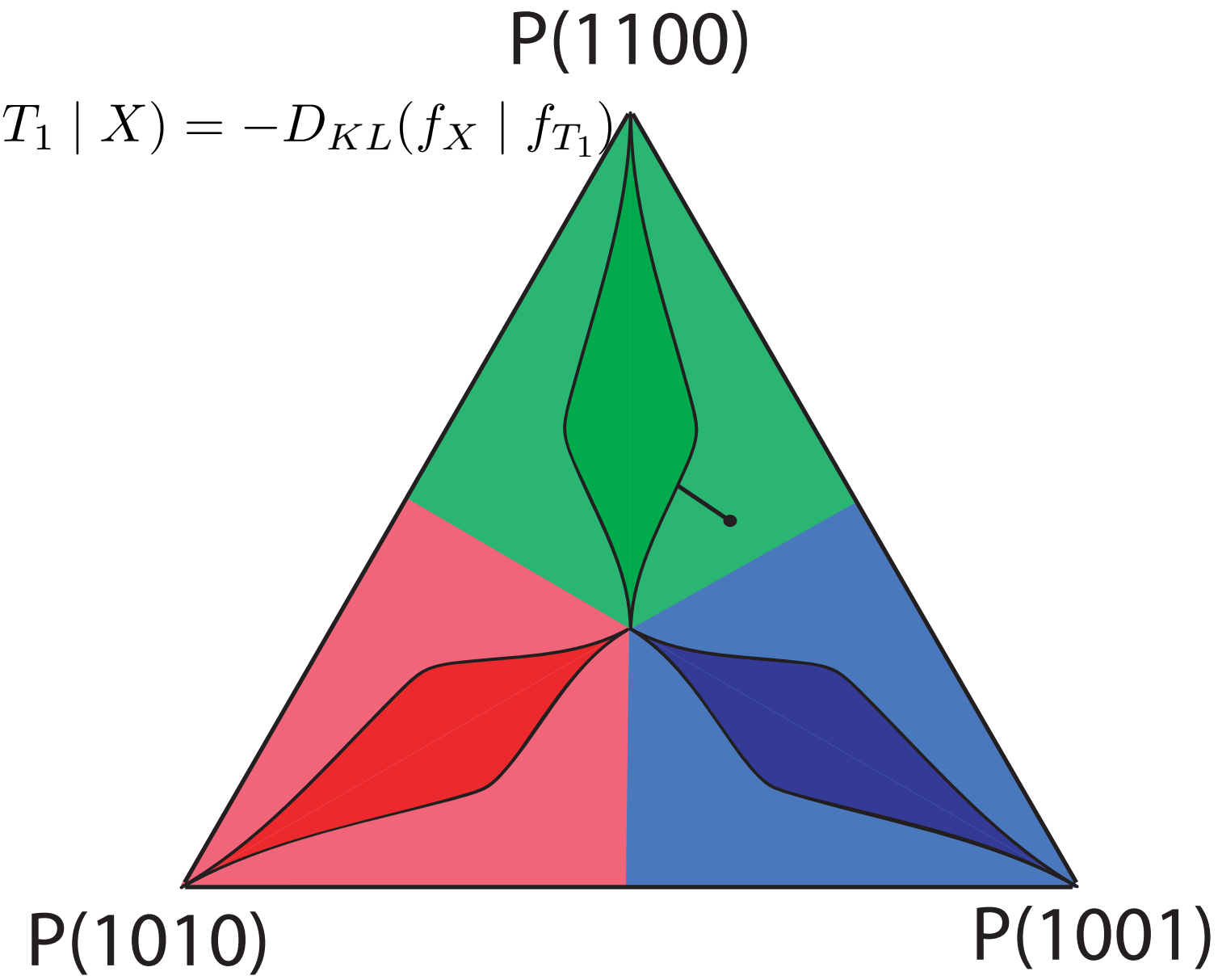


Pattern Frequency Space With Observed Data

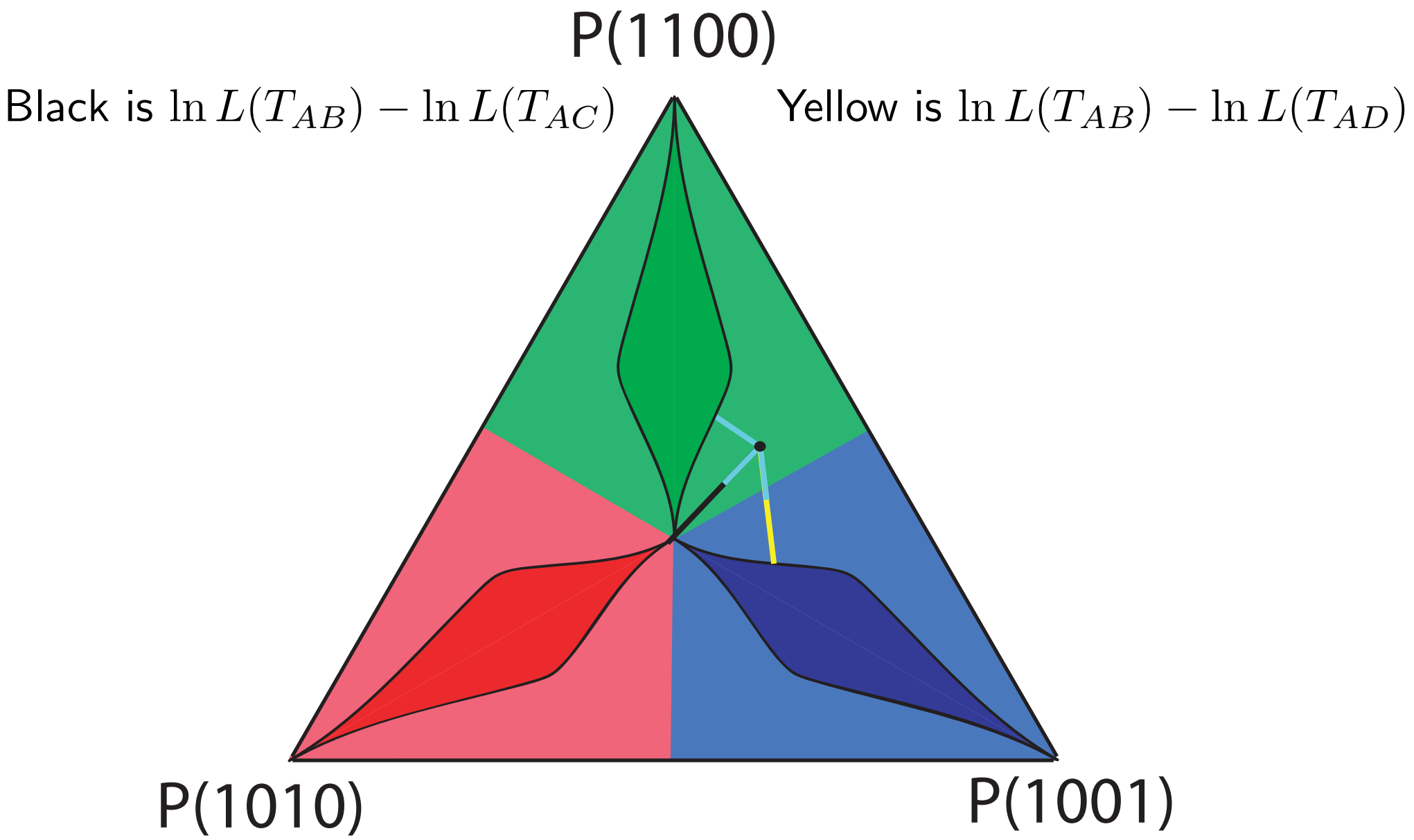


ML scores in Pattern Frequency Space

$$\ln L(T_1 \mid X) = -D_{KL}(f_X \mid f_{T_1})$$

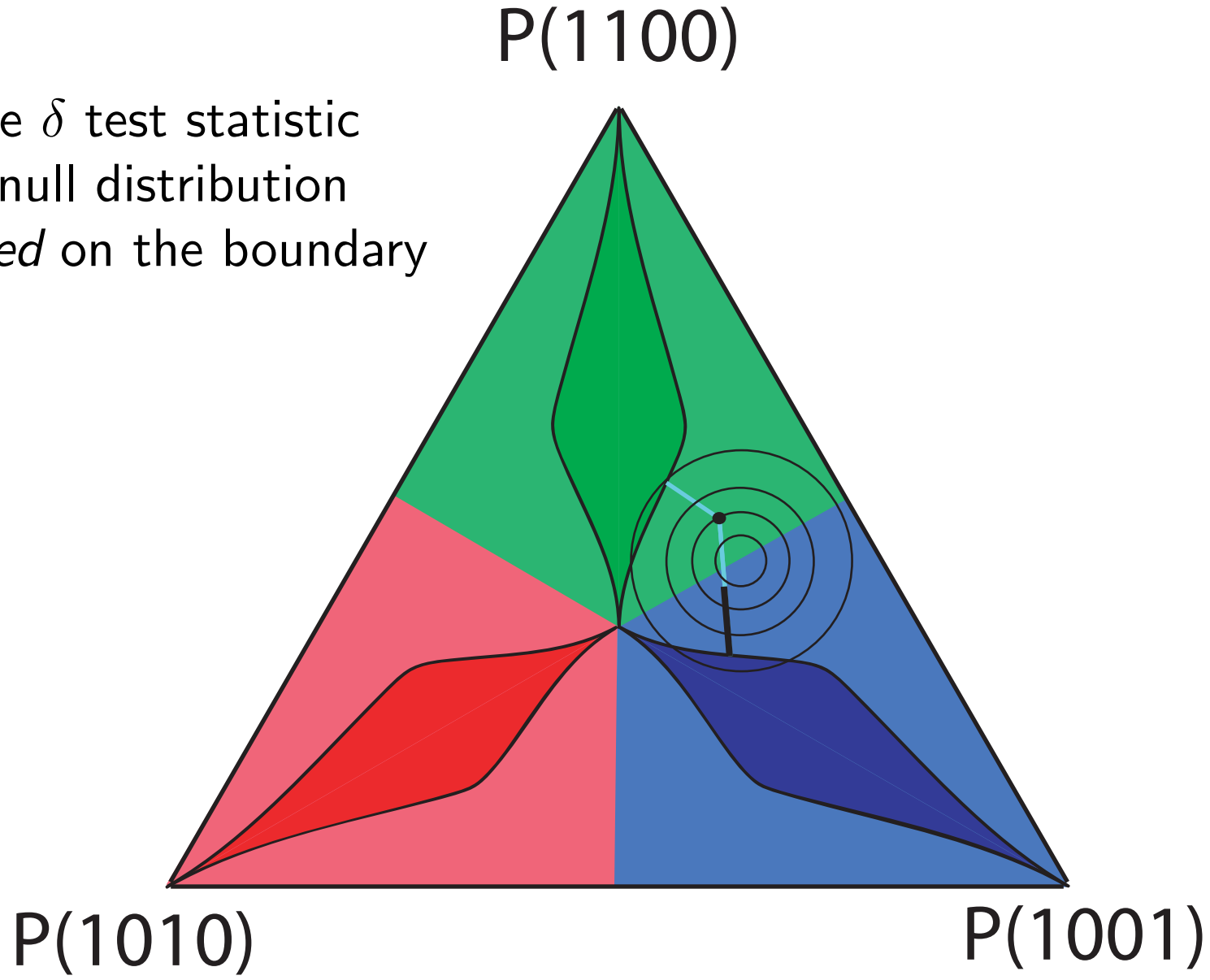


LR statistics in Pattern Frequency Space

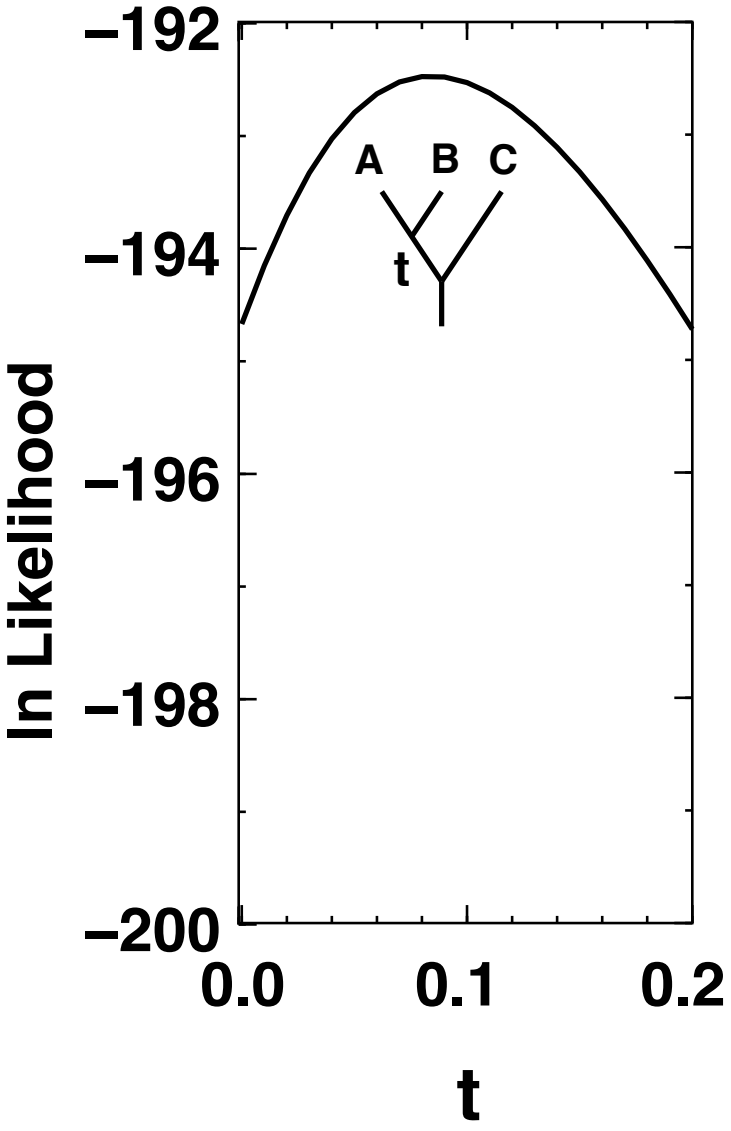


KH Test in Pattern Frequency Space

Uses the δ test statistic
and a null distribution
centered on the boundary



Can we test trees using the LRT?



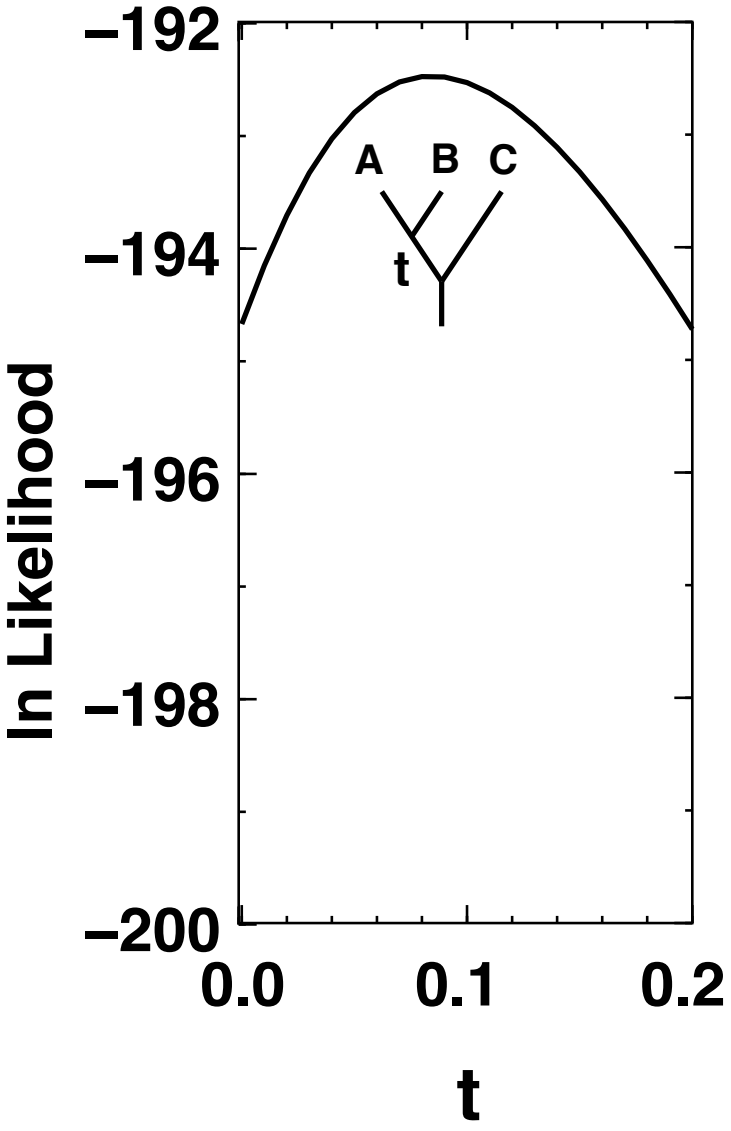
Slide from Joe Felsenstein

1. Should we calculate the LRT as:

$$\delta_i = 2\left[\ln L(t = \hat{t}, T_i \mid X) - \ln L(t = 0, T_i \mid X)\right]$$

2. And can we use the χ^2_1 distribution to get the critical value for δ?

Can we test trees using the LRT?



Slide from Joe Felsenstein

1. Should we calculate the LRT as:

$$\delta_i = 2\left[\ln L(t = \hat{t}, T_i \mid X) - \ln L(t = 0, T_i \mid X)\right]$$

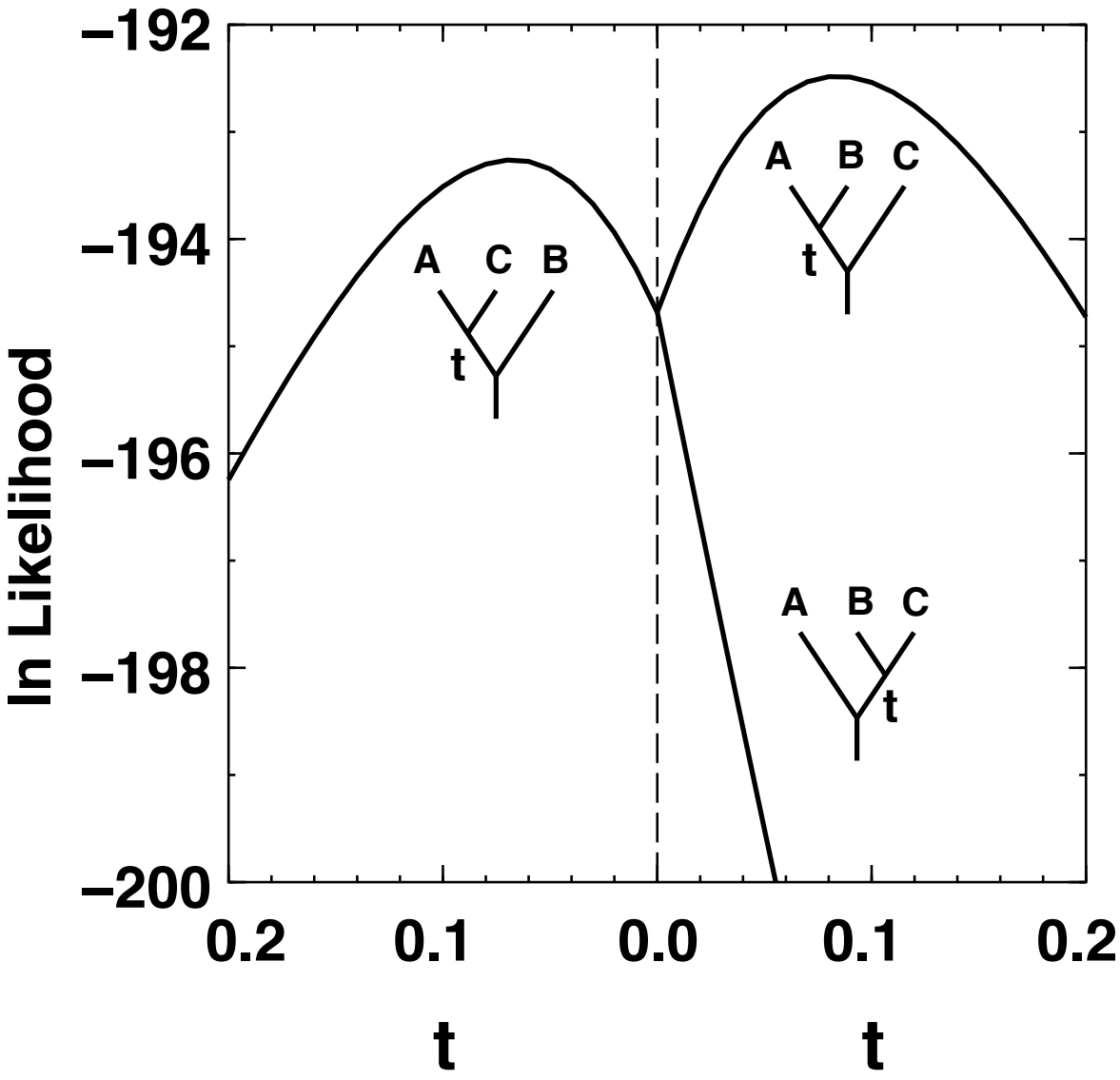
No. $t = 0$ might not yield the best alternative $\ln L$

2. And can we use the χ^2_1 distribution to get the critical value for δ ?

No. Constraining parameters at boundaries leads to a mixture such as: $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$

See Ota et al. (2000).

Can we test trees using the LRT?



No, tree hypotheses are not nested!

Slide from Joe Felsenstein

Other 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.

They are too conservative (Susko, 2014) - more complicated calculations from the Normal [KHns] or mixtures of χ^2 distributions [chi-bar].

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

Summary - Part 1

- $\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$ 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 $\delta(\hat{T}, T_1 \mid X)$

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

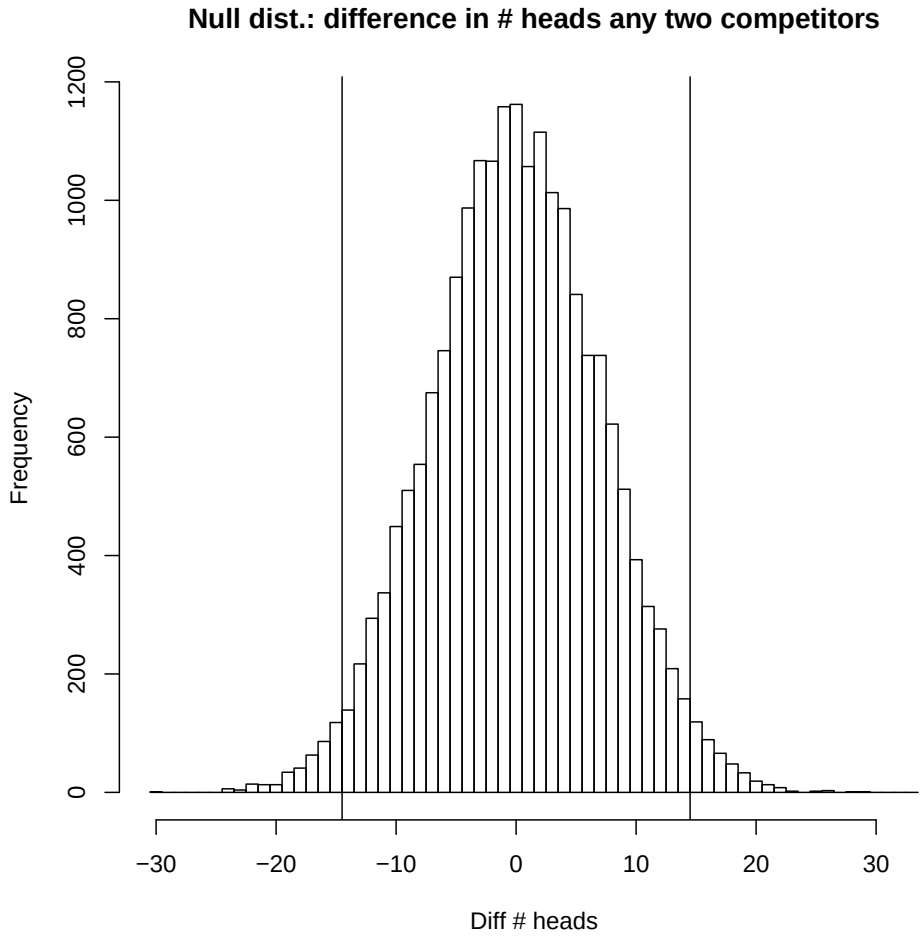
Using the ML tree in your test introduces selection bias

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.

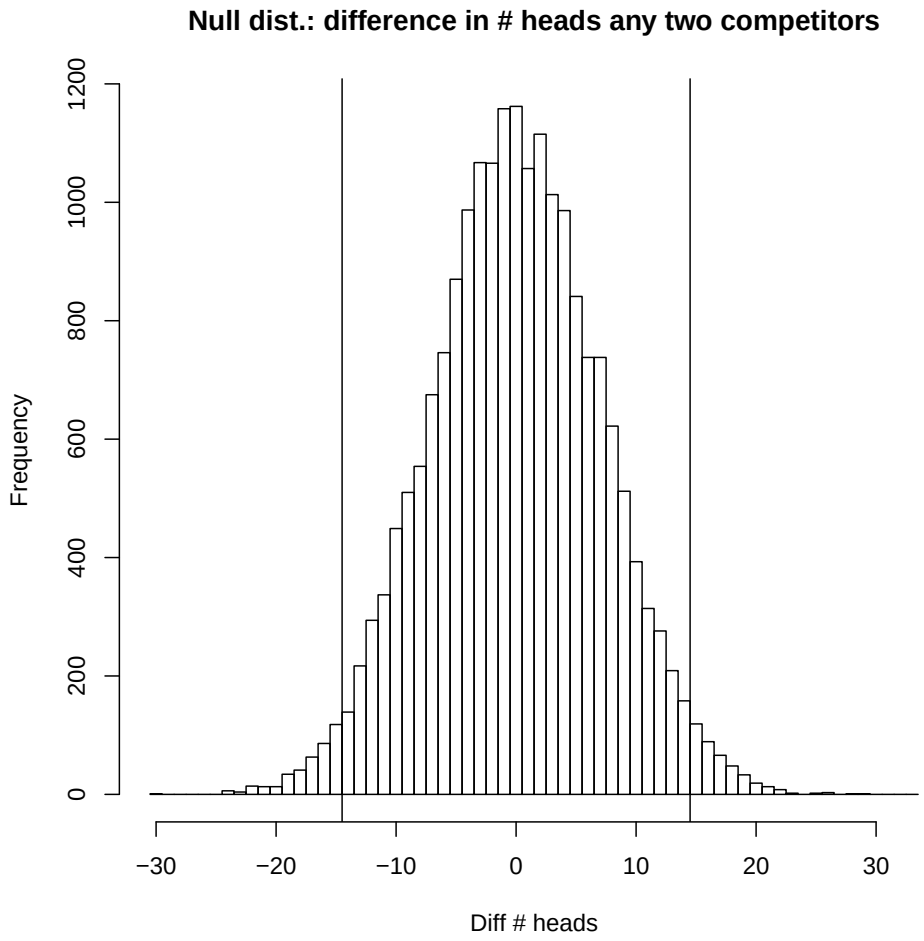
Experiment: 70 people each flip a fair coin 100 times and count # heads.

$$h_1 - h_2$$

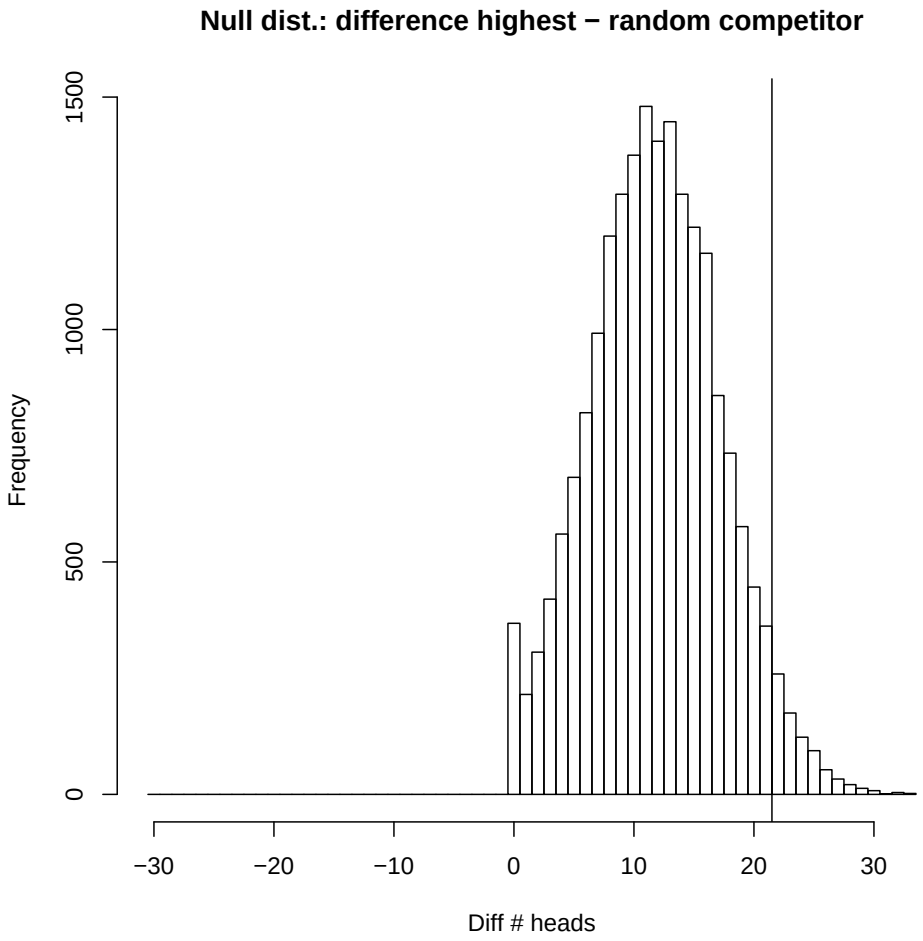


Experiment: 70 people each flip a fair coin 100 times and count # heads.

$h_1 - h_2$



$\max(h) - h_1$



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[2]

[2] Because such a tree would be unlikely to ever be the tree that is the determines the maximum diplacement from the centered value, $m^{(j)}$.

SH Test details

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

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

- 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:

$$\delta_i^{(j)} \geq \delta(\hat{T}, T_i \mid X)$$

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,

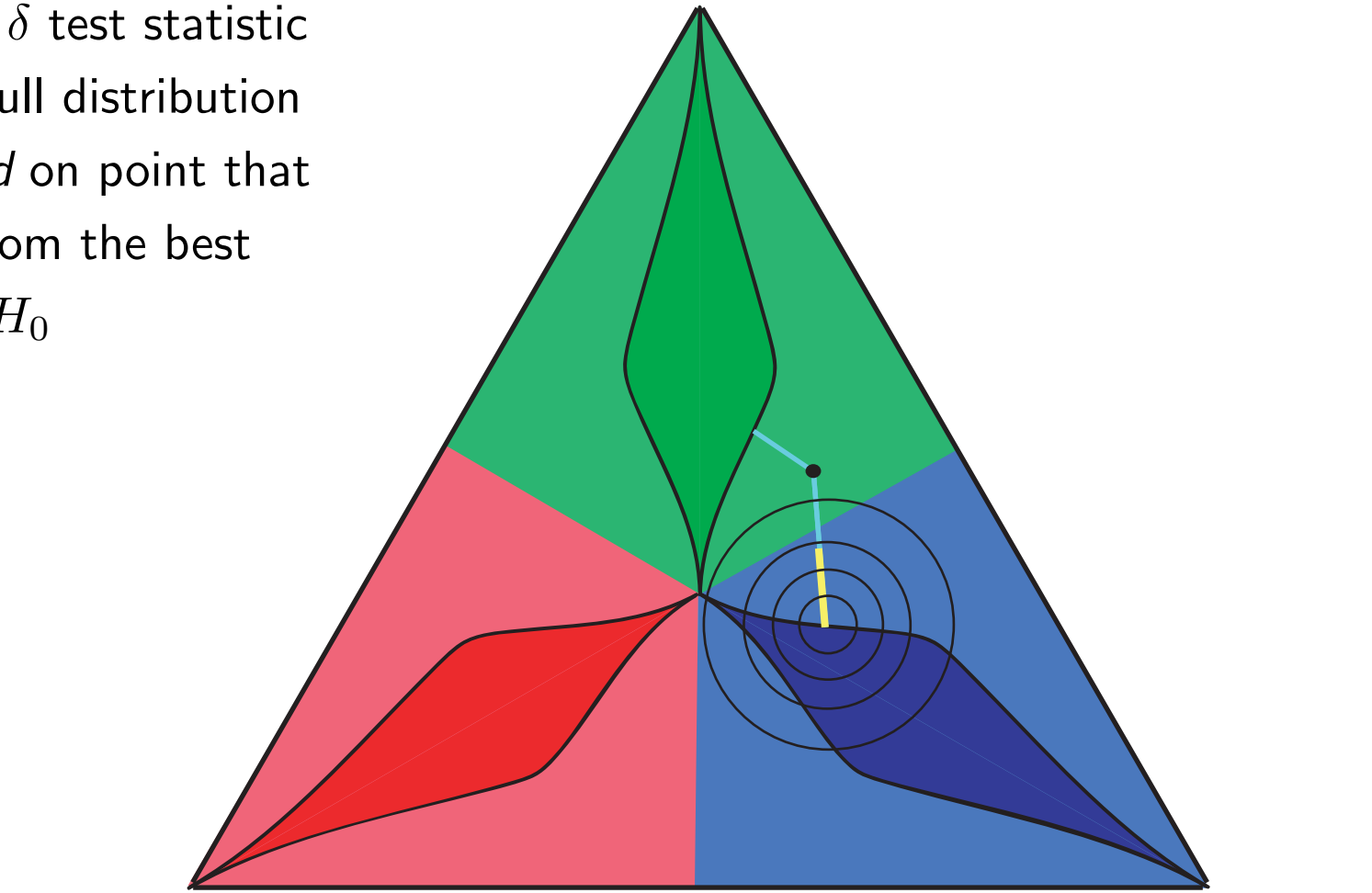
2. Simulate many datasets under the parameters of that model,

3. Calculate $\delta^{(j)} = 2\left[\ln L(\hat{T}^{(j)} \mid X^{(j)}) - \ln L(\hat{T}_0^{(j)} \mid X^{(j)})\right]$ 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

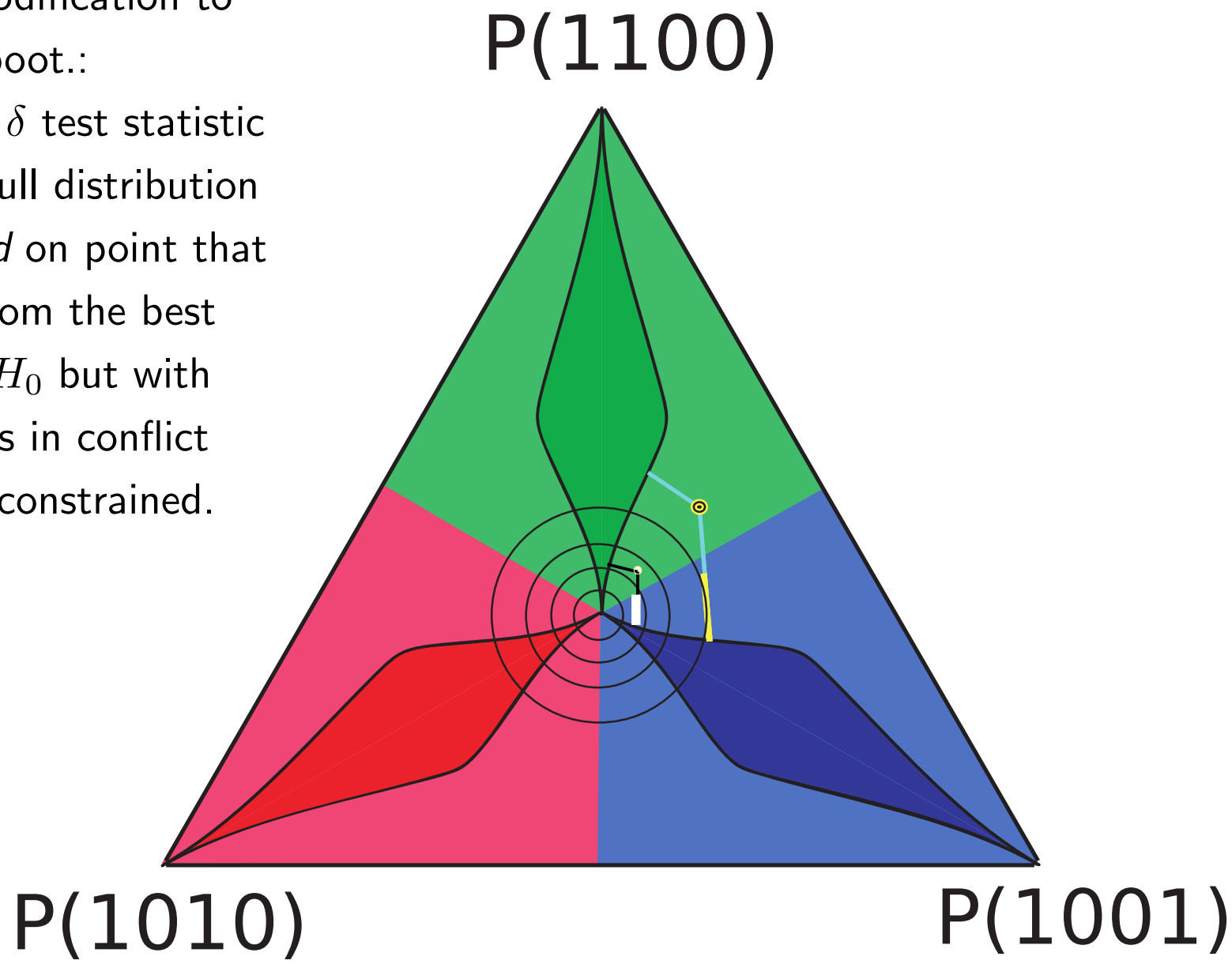
Parametric bootstrapping in Pattern Frequency Space

Uses the δ test statistic and a null distribution *centered* on point that arises from the best tree in H_0



Susko modification to param. boot.:

Uses the δ test statistic and a null distribution *centered* on point that arises from the best tree in H_0 but with branches in conflict with $\hat{T}$ constrained. to be 0.



Parametric bootstrapping

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

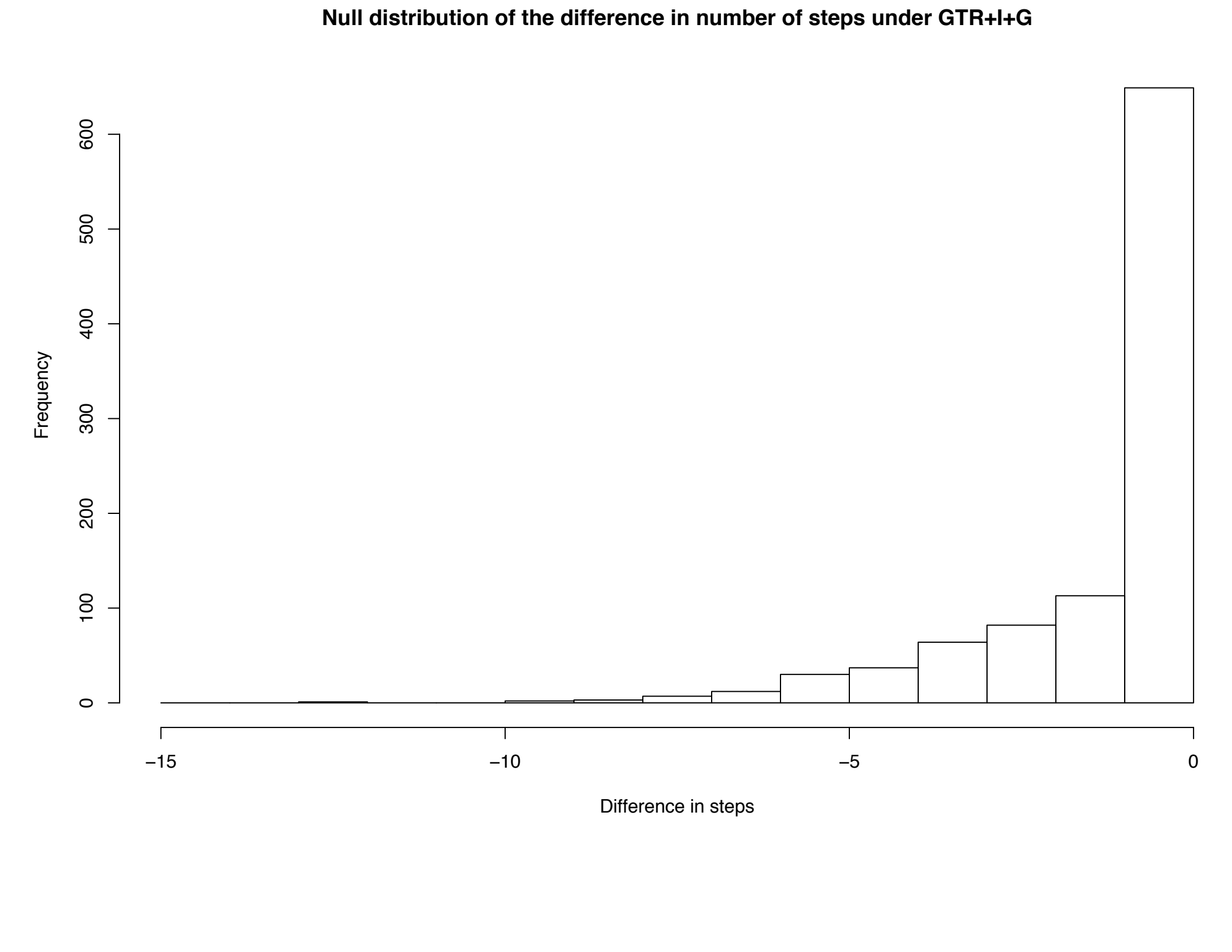
Huelsenbeck et al. (1996) describes how to use the approach as a test for monophyly.

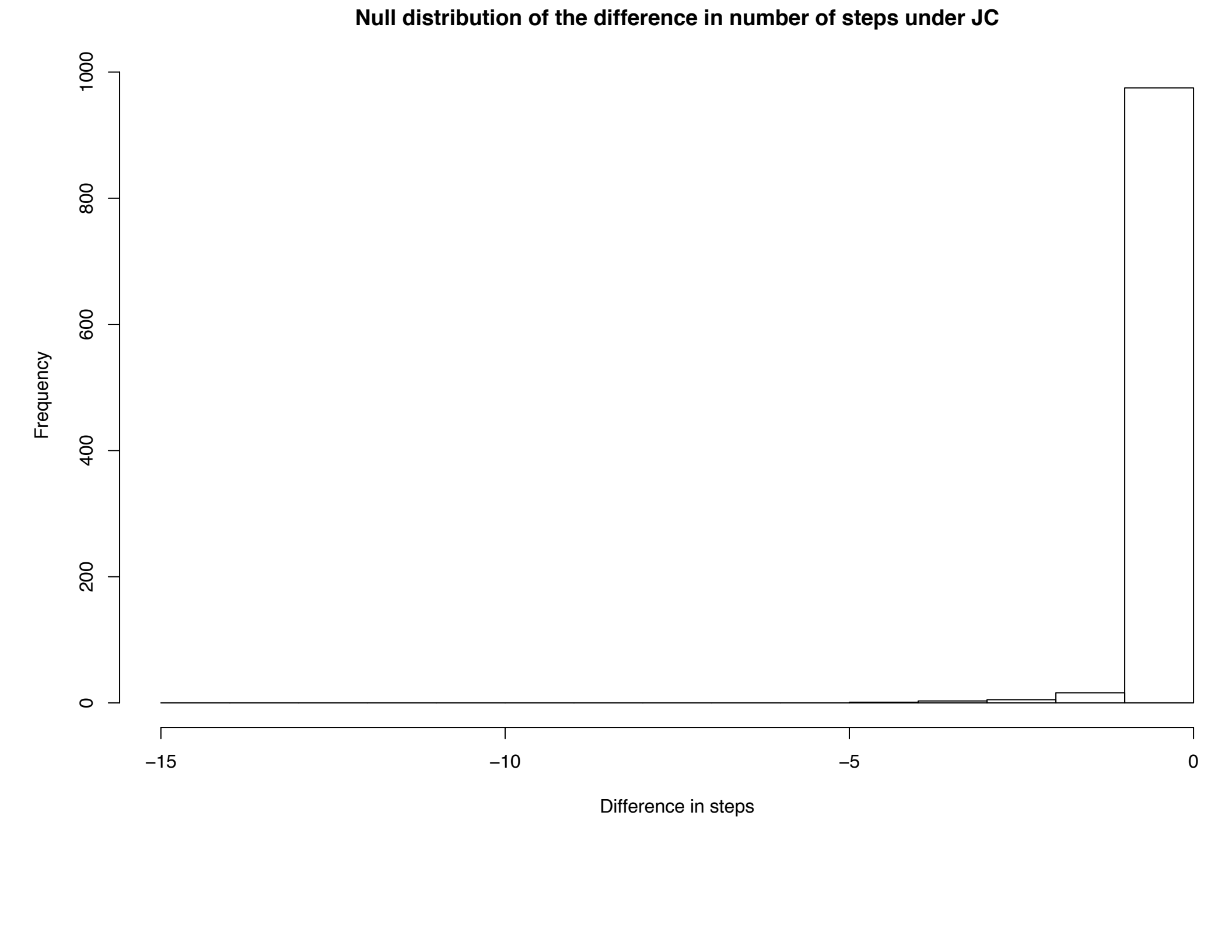
Intuitive and powerful, but not robust to model violation (Buckley, 2002).

Can be done manually[3] or via SOWHAT (demo tonight)

Susko (2014): collapse optimize null tree with 0-length contraints for the branch in question (to avoid rejecting too often)

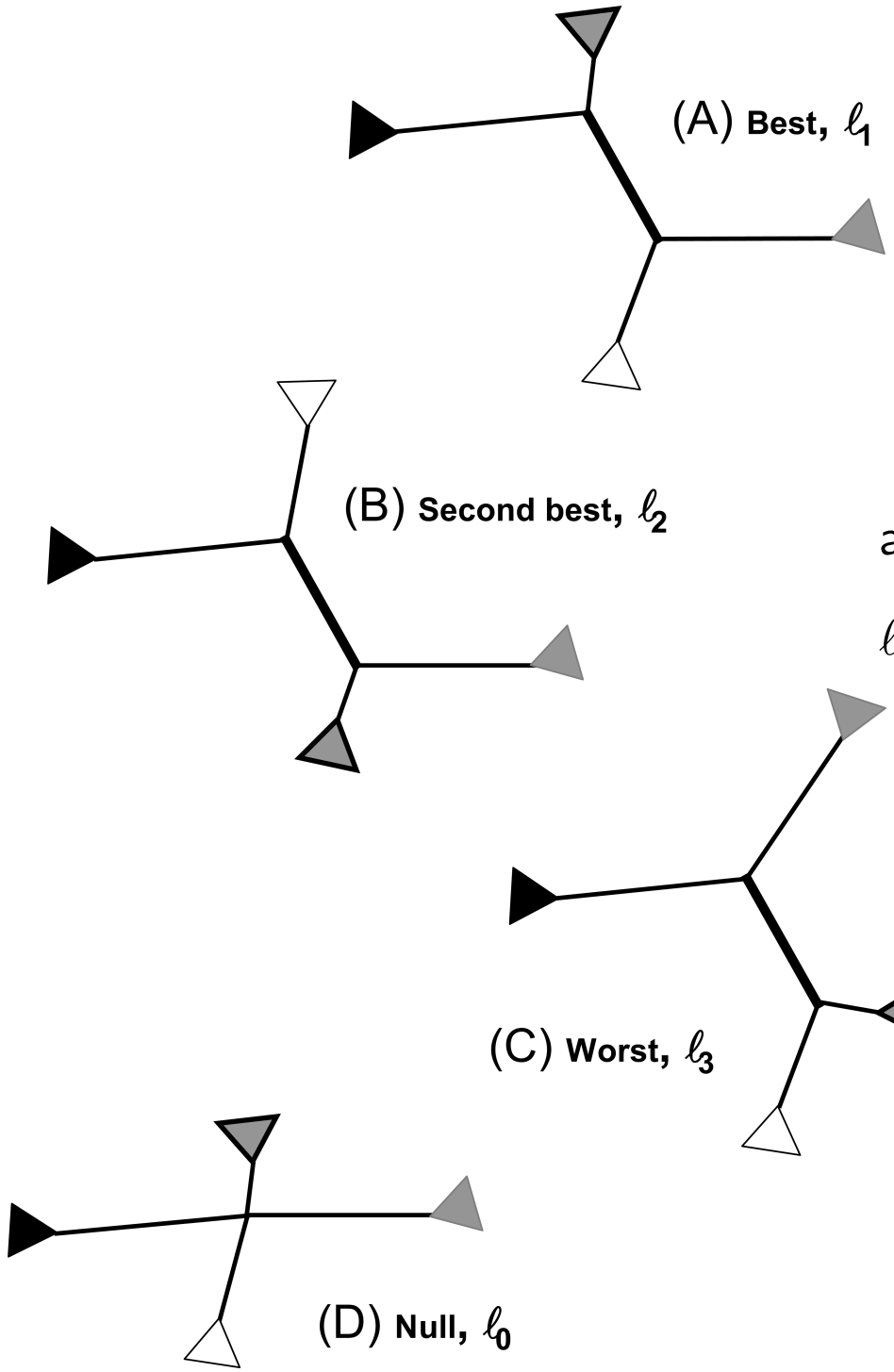
[3] instructions in https://molevol.mbl.edu/wiki/index.php/ParametricBootstrappingLab


Null distribution of the difference in number of steps under GTR+I+G
Frequency
0
100
200
300
400
500
600
−15
−10
−5
0
Difference in steps


Null distribution of the difference in number of steps under JC
Frequency
0
200
400
600
800
1000
−15
−10
−5
0
Difference in steps

aLRT of Anisimova and Gascuel (2006)

- For a **branch** j, calculate $\delta_j^\dagger$ as twice the difference in $\ln L$ between the optimal tree (which has the branch) and the best NNI neighbor.
- This is very fast.
- They argue that the null distribution for each LRT around the polytomy follows a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution
- The introduce Bonferroni-correction appropriate for correcting for the selection of the best of the three resolutions.
- They find aLRT to be accurate and powerful in simulations, but Anisimova et al. (2011) report that it rejects too often and is sensitive to model violation.



$$\text{aLRT} = 2\left[\ln \ell_1 - \ln L(T_2 \mid X)\right]$$

$$\ell_1 = L(T_1 \mid X)$$

Image from Anisimova and Gascuel (2006)

aBayes Anisimova et al. (2011)

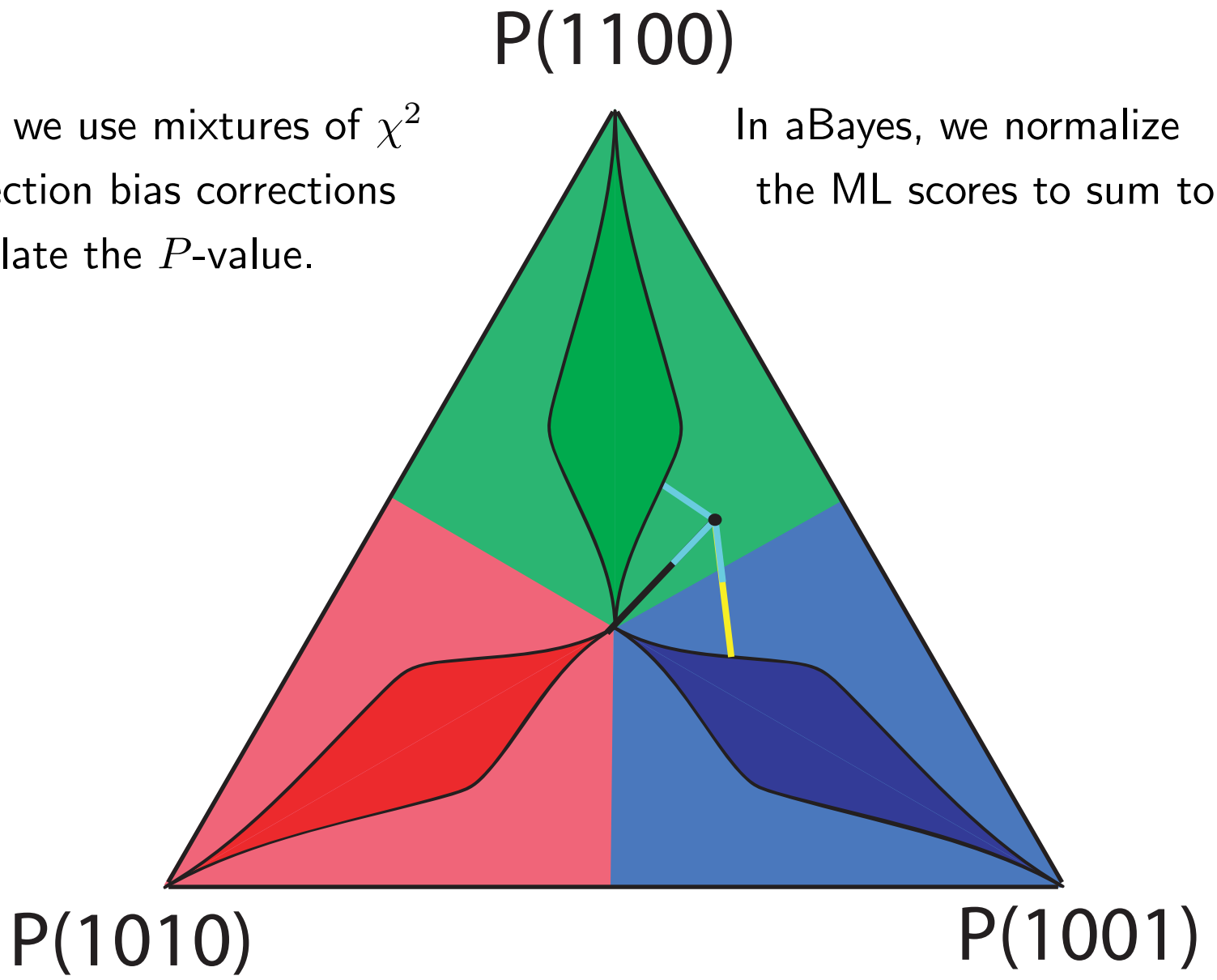
$$\text{aBayes}(T_1 \mid X) = \frac{\Pr(X \mid T_1)}{\Pr(X \mid T_1) + \Pr(X \mid T_2) + \Pr(X \mid T_3)}$$

Simulation studies of Anisimova et al. (2011) show it to have the best power of the methods that do not have inflated probability of falsely rejecting the null.

It is sensitive to model violation.

This is similar to "likelihood-mapping" of Strimmer and von Haeseler (1997)

aLRT and aBayes in Pattern Frequency Space



In aLRT, we use mixtures of χ^2 and selection bias corrections to calculate the P-value.

In aBayes, we normalize the ML scores to sum to 1

Bootstrap proportions have been characterized as providing:

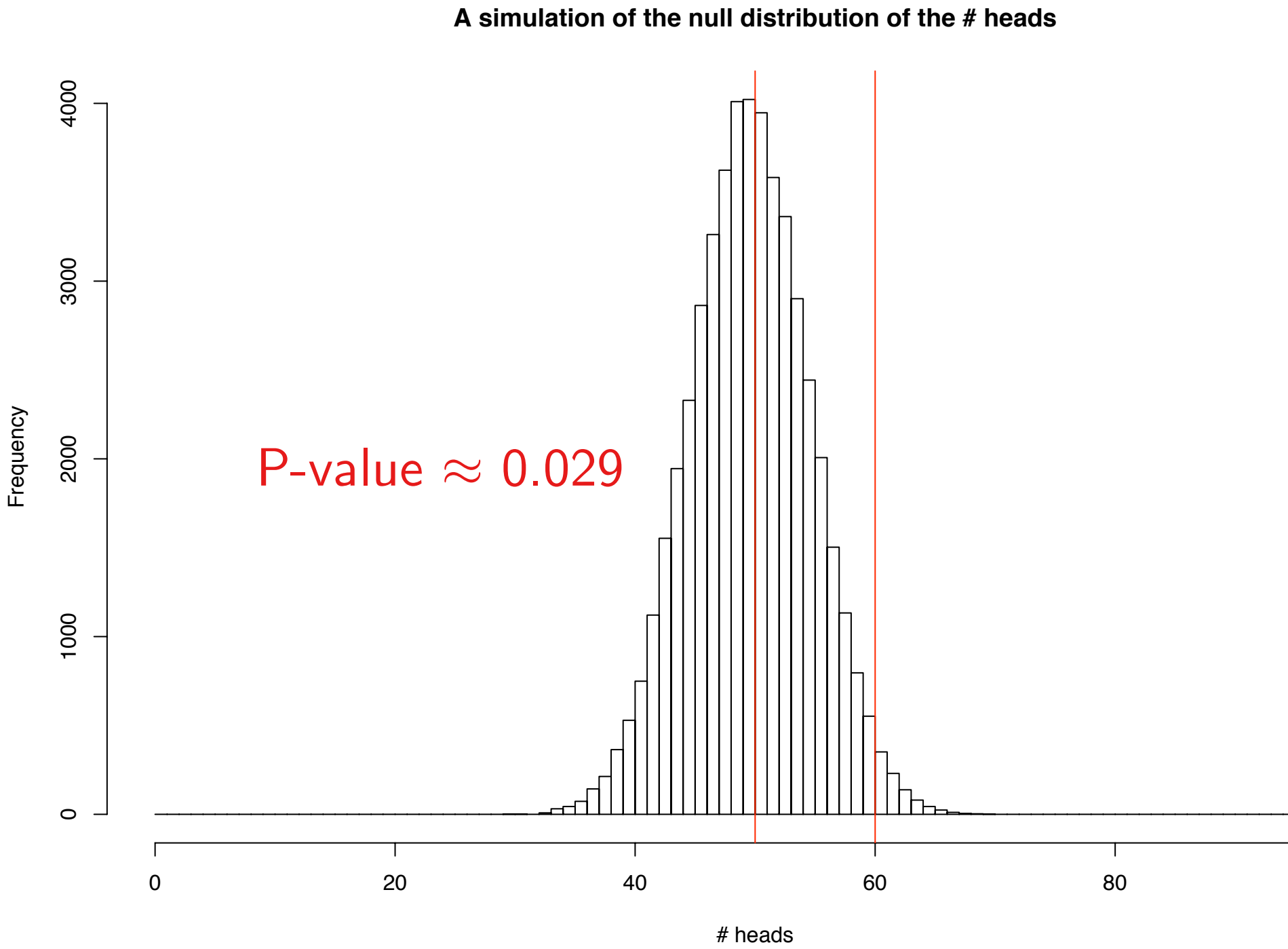
- a measure of repeatability,
- an estimate of the probability that the tree is correct (and bootstrapping has been criticized as being too conservative in this context),
- the P-value for a tree or clade

coin flipping (yet again)

$N = 100$ and $H = 60$

Can we reject the hypothesis of a fair coin?

We can use simulation to generate the null distribution (we could actually use the binomial distribution to analytically solve this one)...


A simulation of the null distribution of the # heads
Frequency
0
1000
2000
3000
4000
P-value ≈ 0.029
0
20
40
60
80
heads

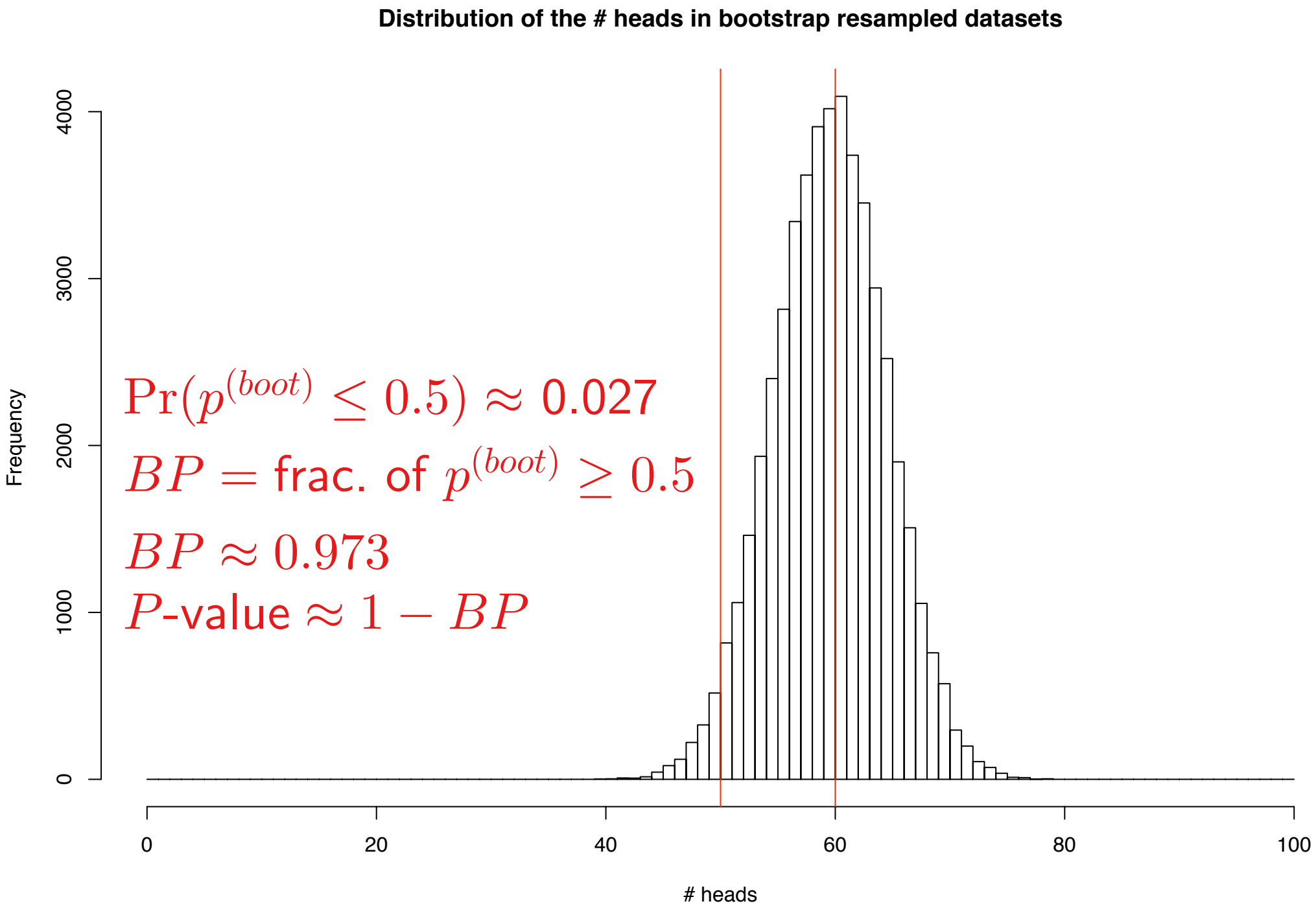
We discussed how bootstrapping gives us a sense of the variability ofour estimate

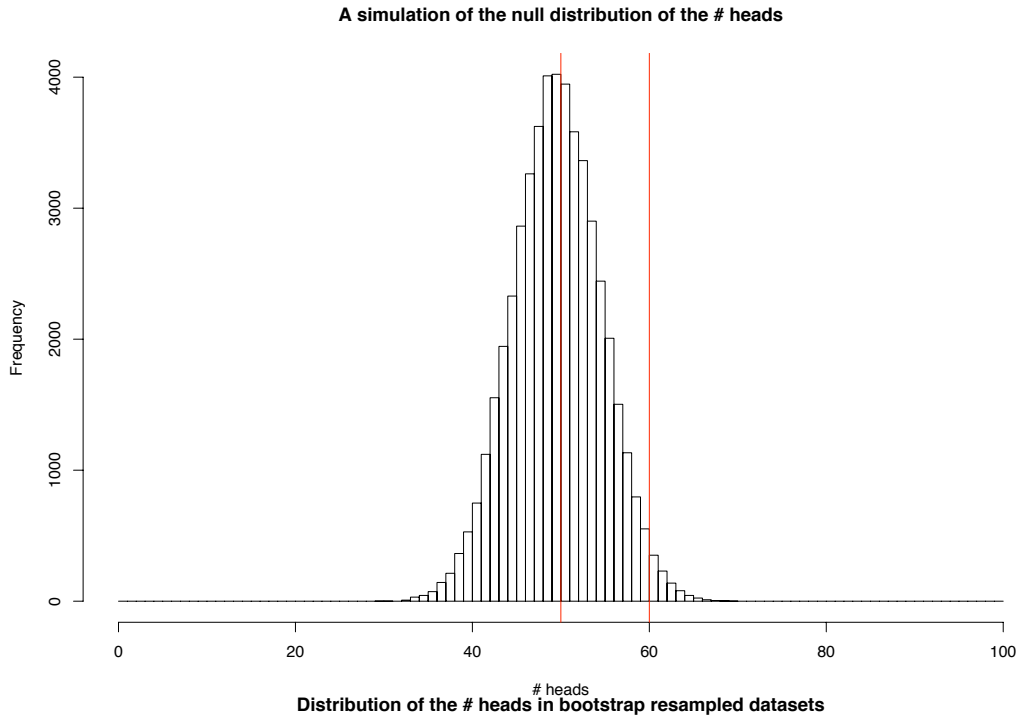
It can also give a tail probability for $\Pr(f_H^{(boot)} \leq 0.5)$

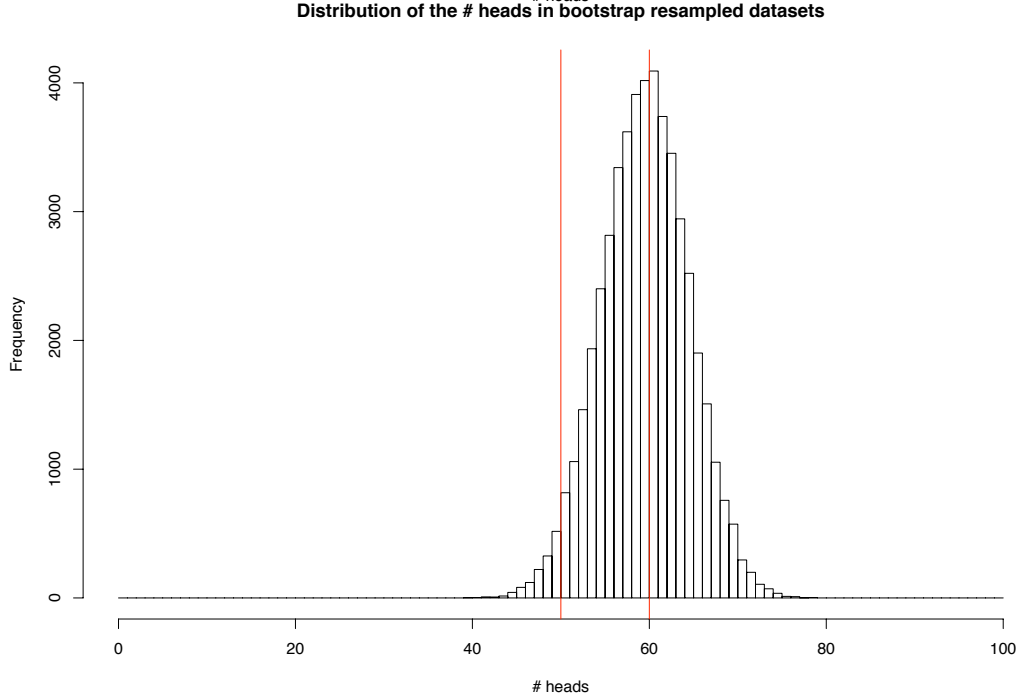
Amazingly (for many applications):

$$\Pr(\hat{f}_H \geq 0.6 \mid \text{null is true}) \approx \Pr(f_H^{(boot)} \leq 0.5)$$

In other words, the P-value is approximate by the fraction of bootstrap replicates consistent with the null.


Distribution of the # heads in bootstrap resampled datasets
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
heads
$\Pr(p^{(boot)} \leq 0.5) \approx 0.027$
$BP =$ frac. of $p^{(boot)} \geq 0.5$
$BP \approx 0.973$
P-value $\approx 1 - BP$


A simulation of the null distribution of the # heads
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
heads

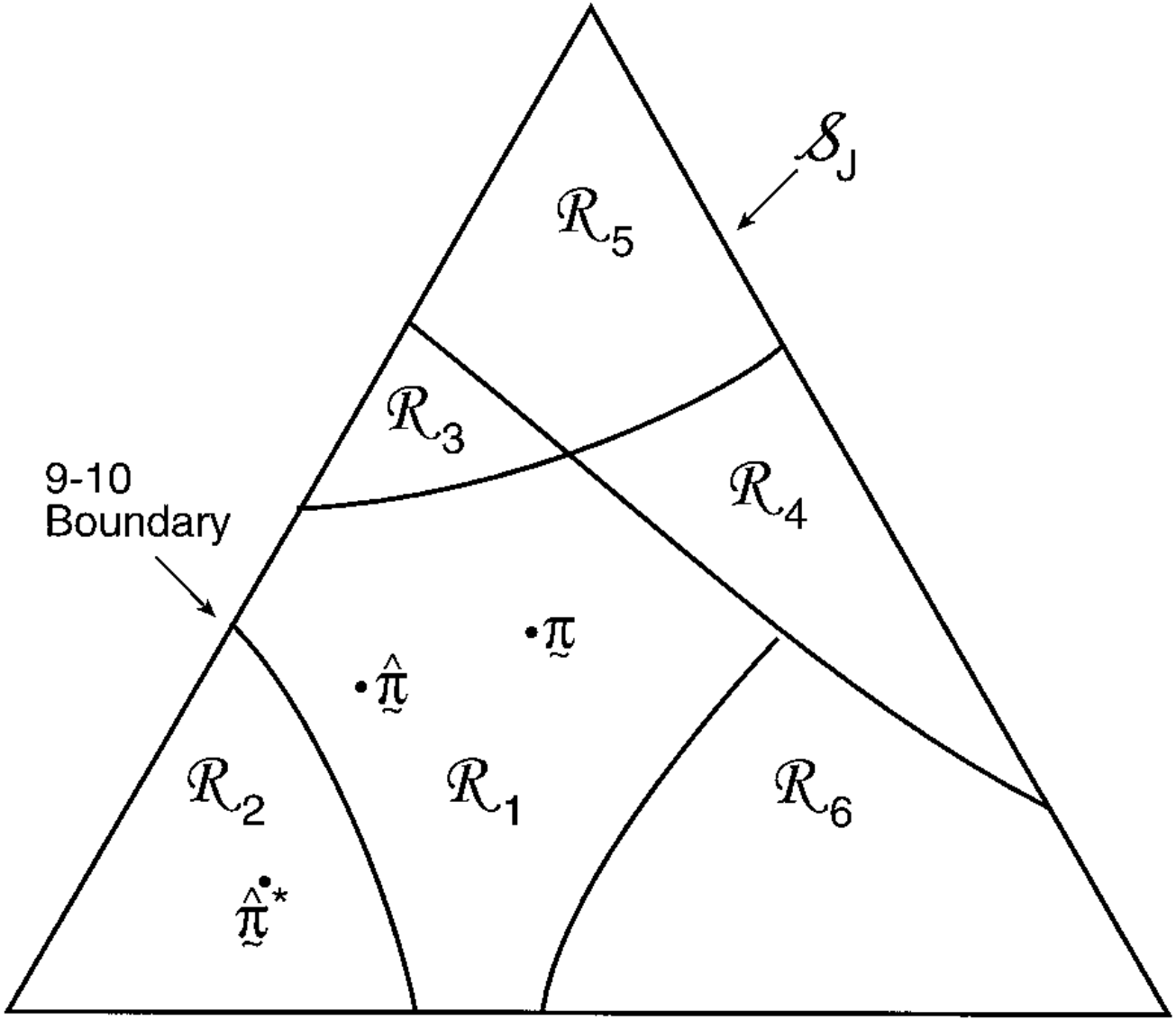

Distribution of the # heads in bootstrap resampled datasets
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
heads

- When you decide between trees, the boundaries between tree hypotheses can be curved

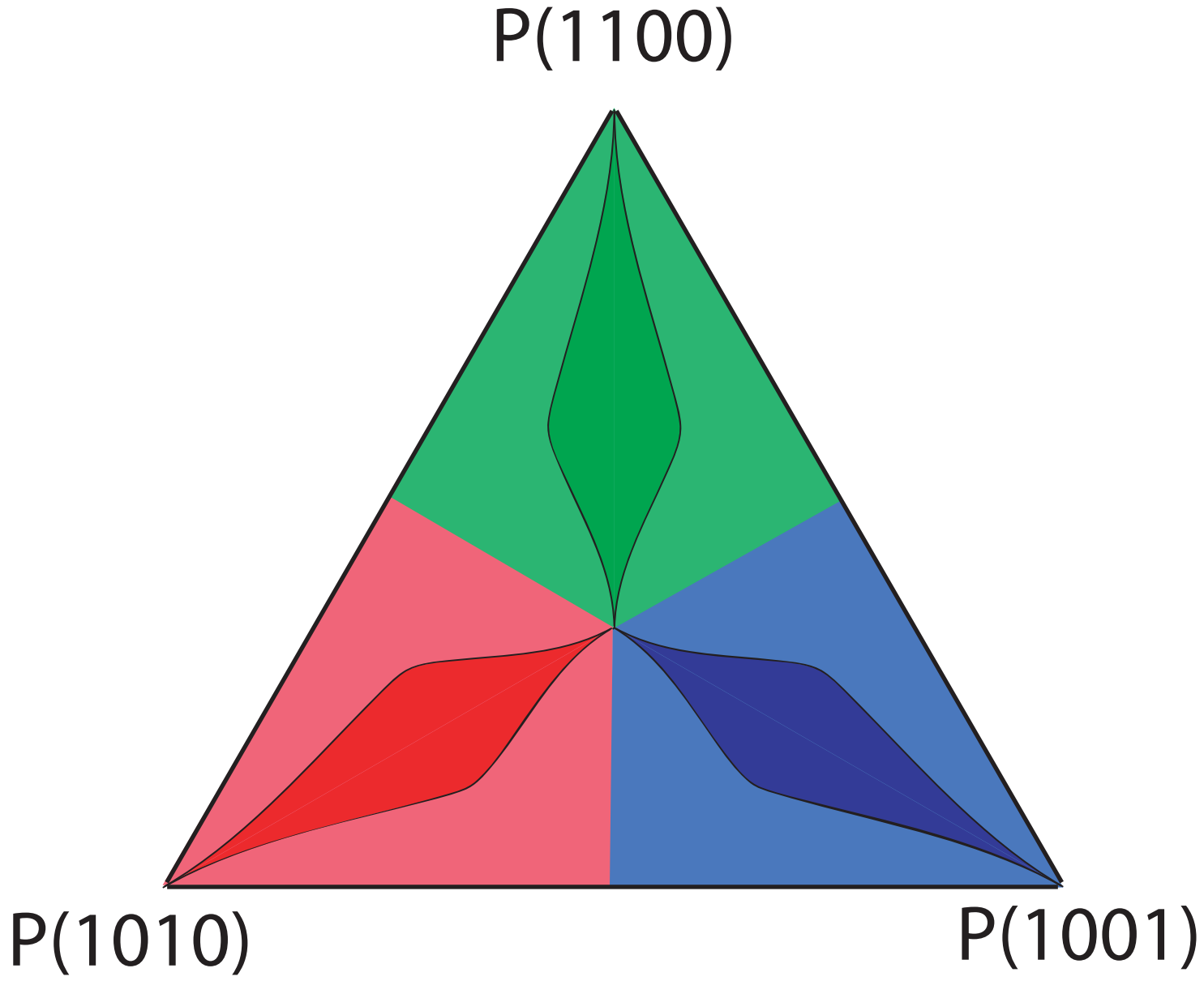
- When the boundary of the hypothesis space is curved, 1 - BP can be a poor approximation of the P-value.

– Efron et al. (1996)

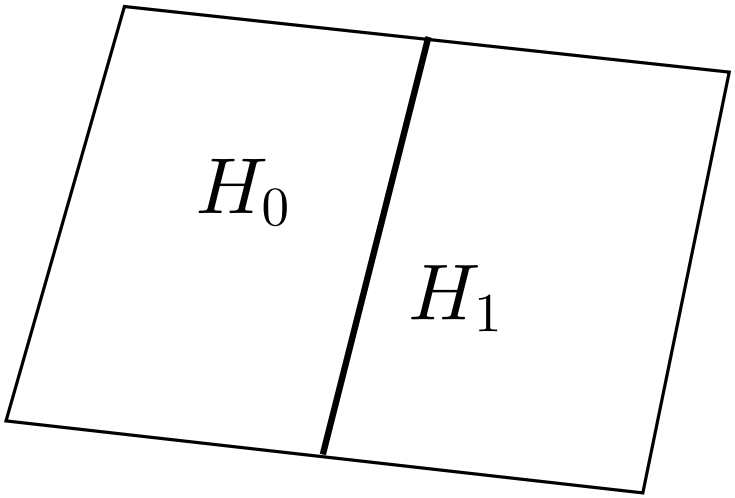
Efron et al. (1996) view of tree space

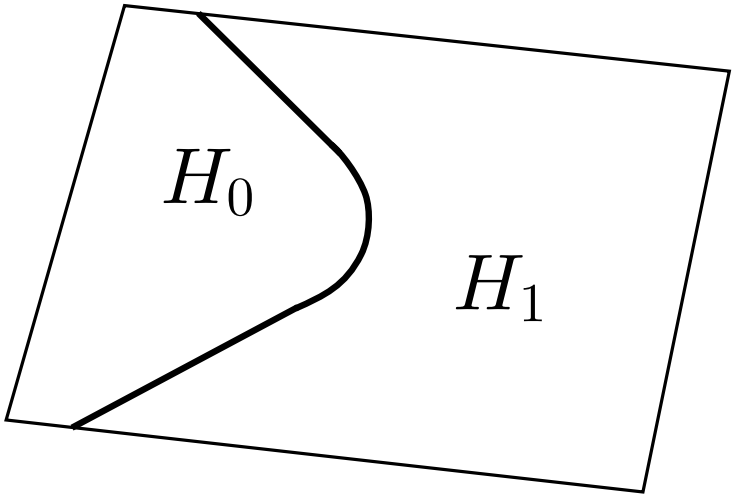


Parsimony-informative Pattern Frequency Space



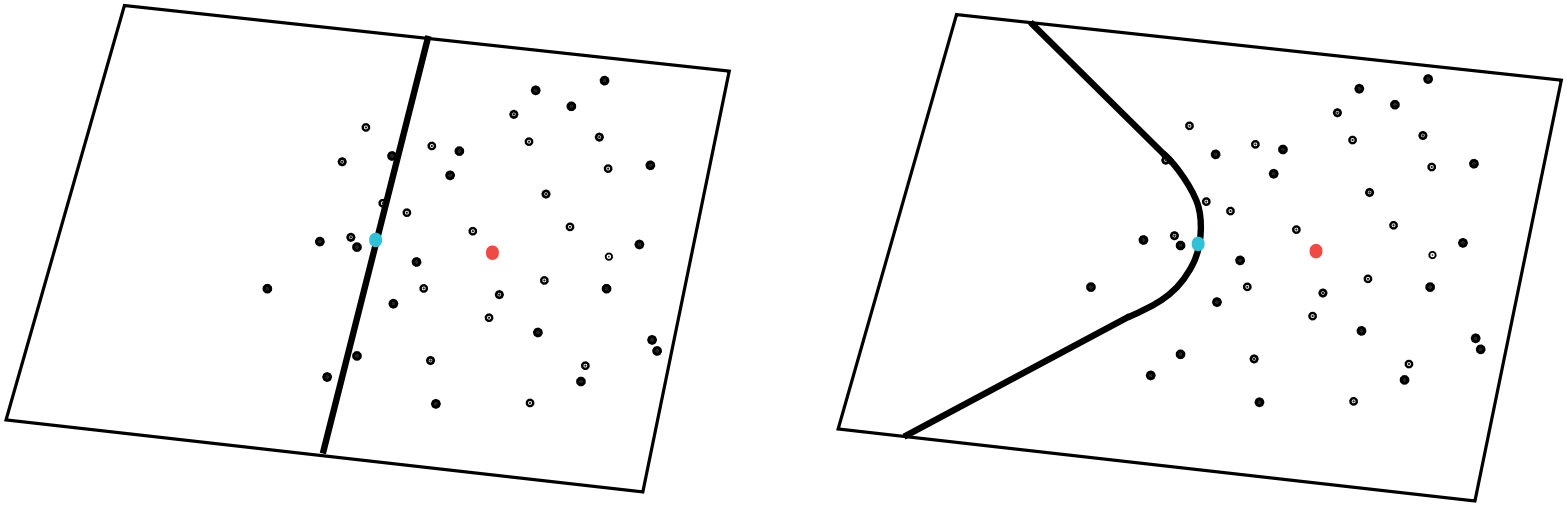
Imagine hypothesis tests of locations with different border shapes:





Similar dataset with point estimates (red dot) in H_1

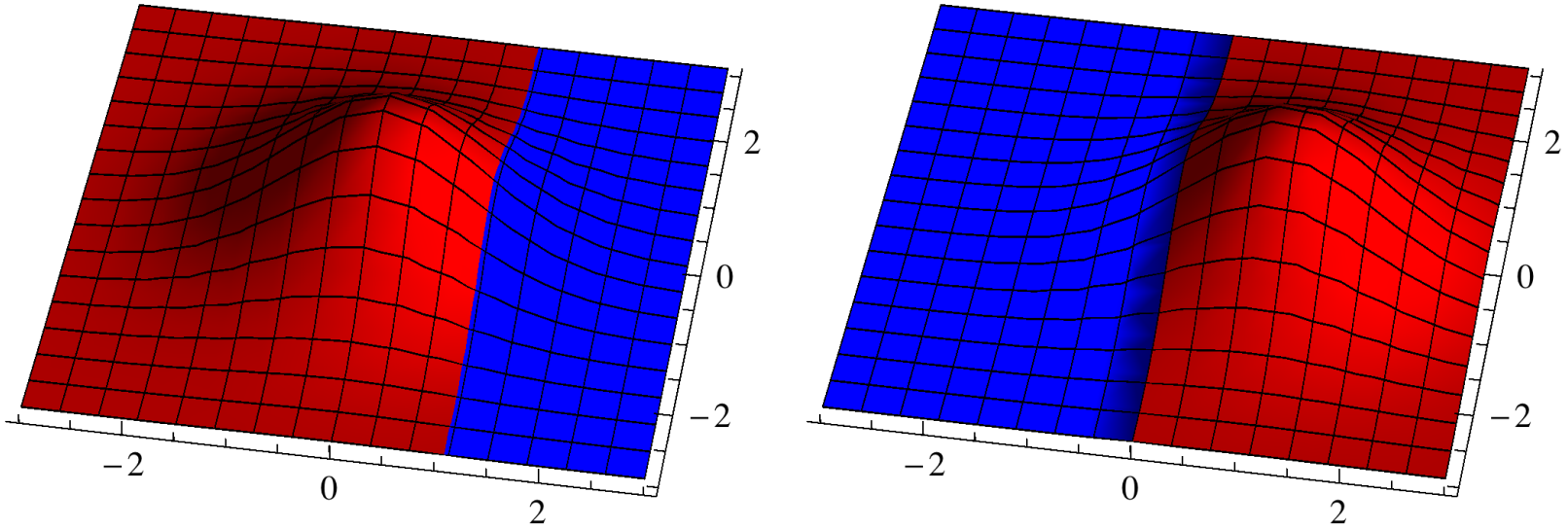
Green dot is the hardest set of locations in H_0 to reject.



In the straight border case, symmetry implies that:

The actual P-value (blue region) $\approx 1 - BP$

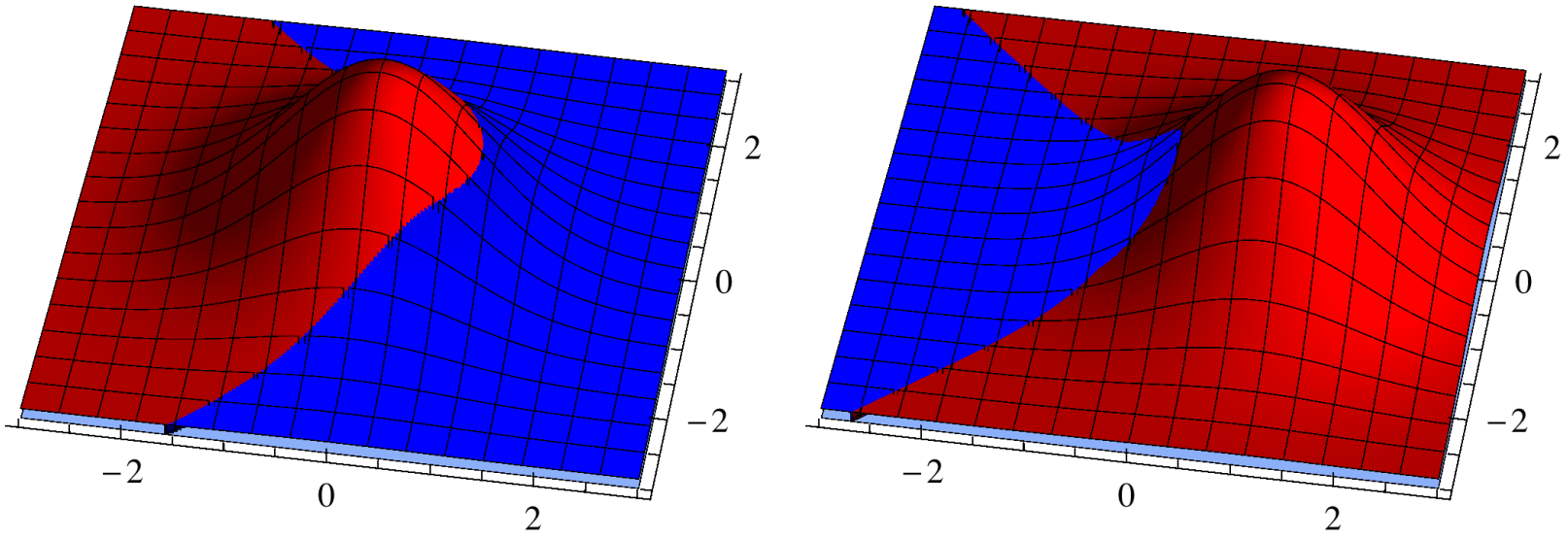
($1 - BP$ is the blue below)



In the curved border case, the symmetry breaks down:

The actual P-value (blue region) $\neq 1 - BP$

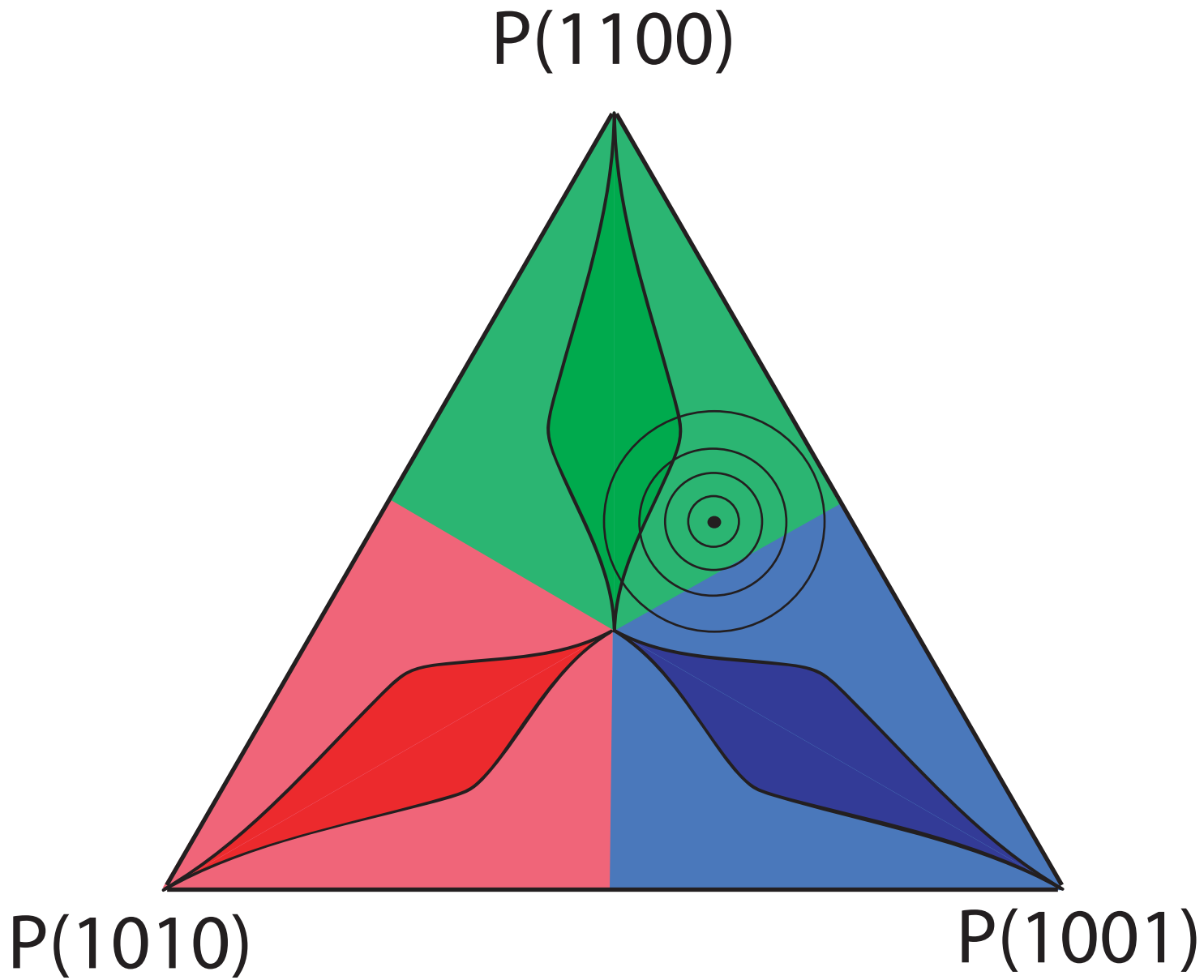
($1 - BP$ is the blue below)



- Efron et al. (1996) proposed a computationally expensive multi-level bootstrap (which has not been widely used).

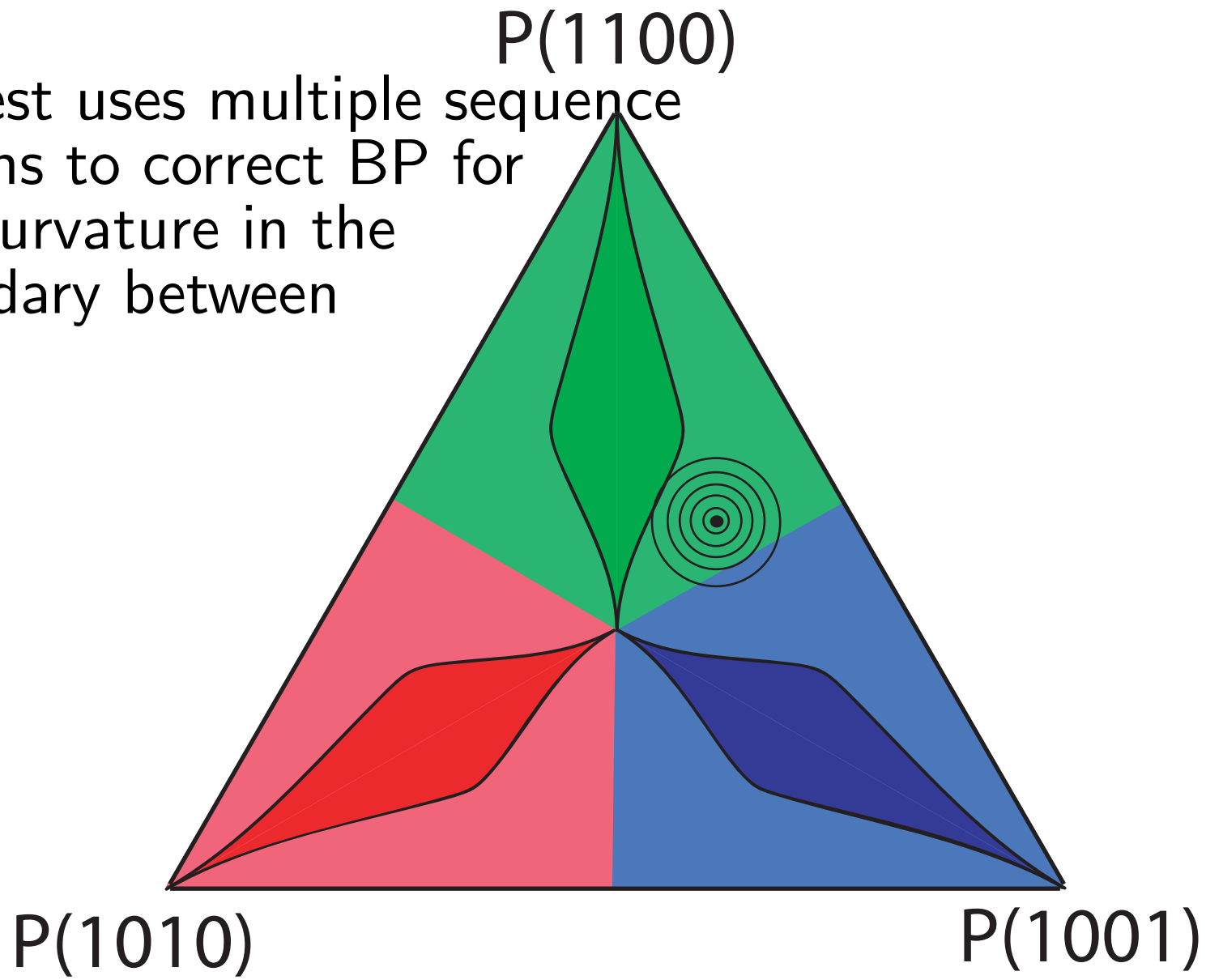
- Shimodaira (2002) used the same theoretical framework to devise a (more feasible) Approximately Unbiased (AU) test of topologies.
  - Multiple scales of bootstrap resampling (80% of characters, 90%, 100%, 110%...) are used to detect and correct for curvature of the boundary.
  - Implemented in the new versions of PAUP*

Non-parametric Bootstrapping in Pattern Frequency Space



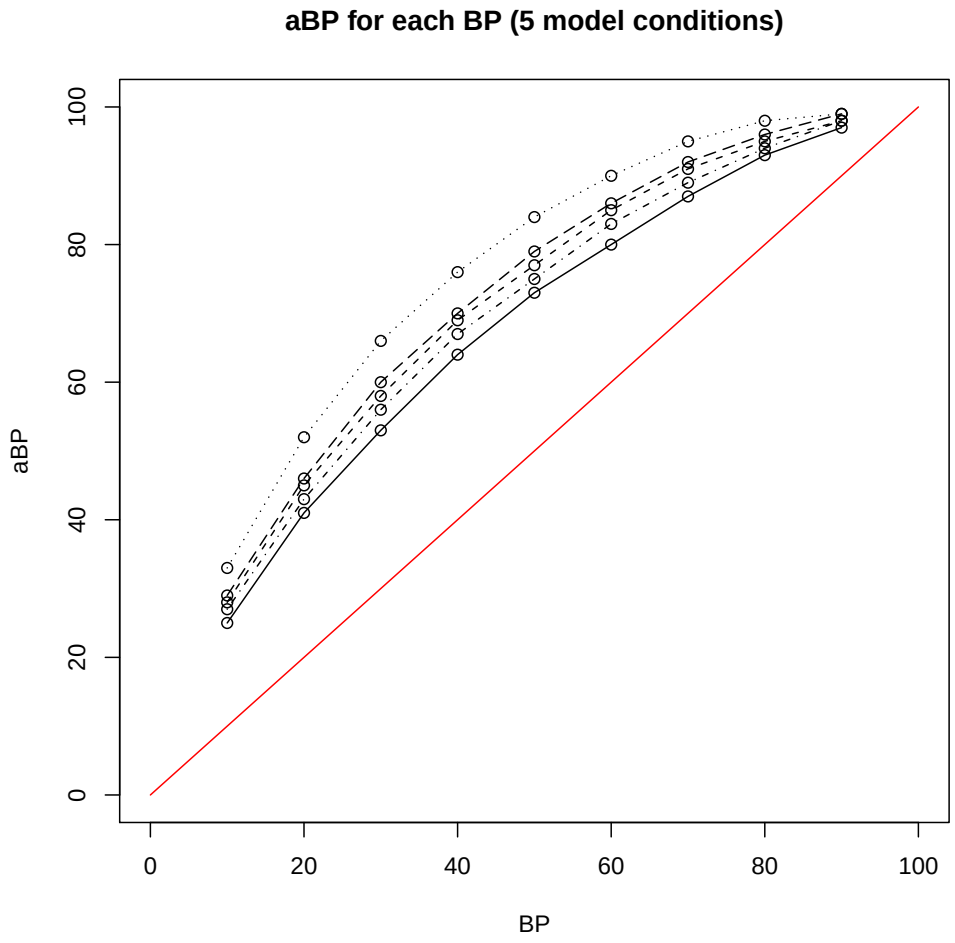
Bootstrapping in Pattern Frequency Space (if you had more data)

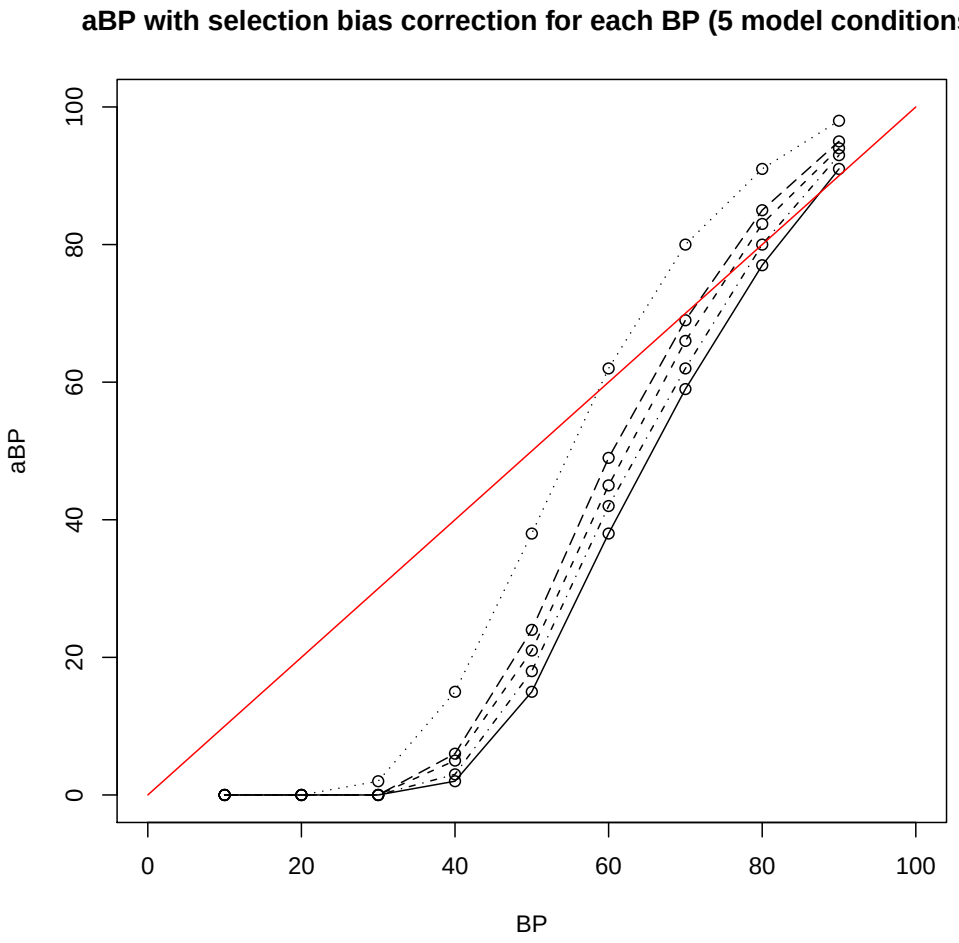
AU Test uses multiple sequence lengths to correct BP for any curvature in the boundary between trees



Susko (2010) adjusted BP – aBP

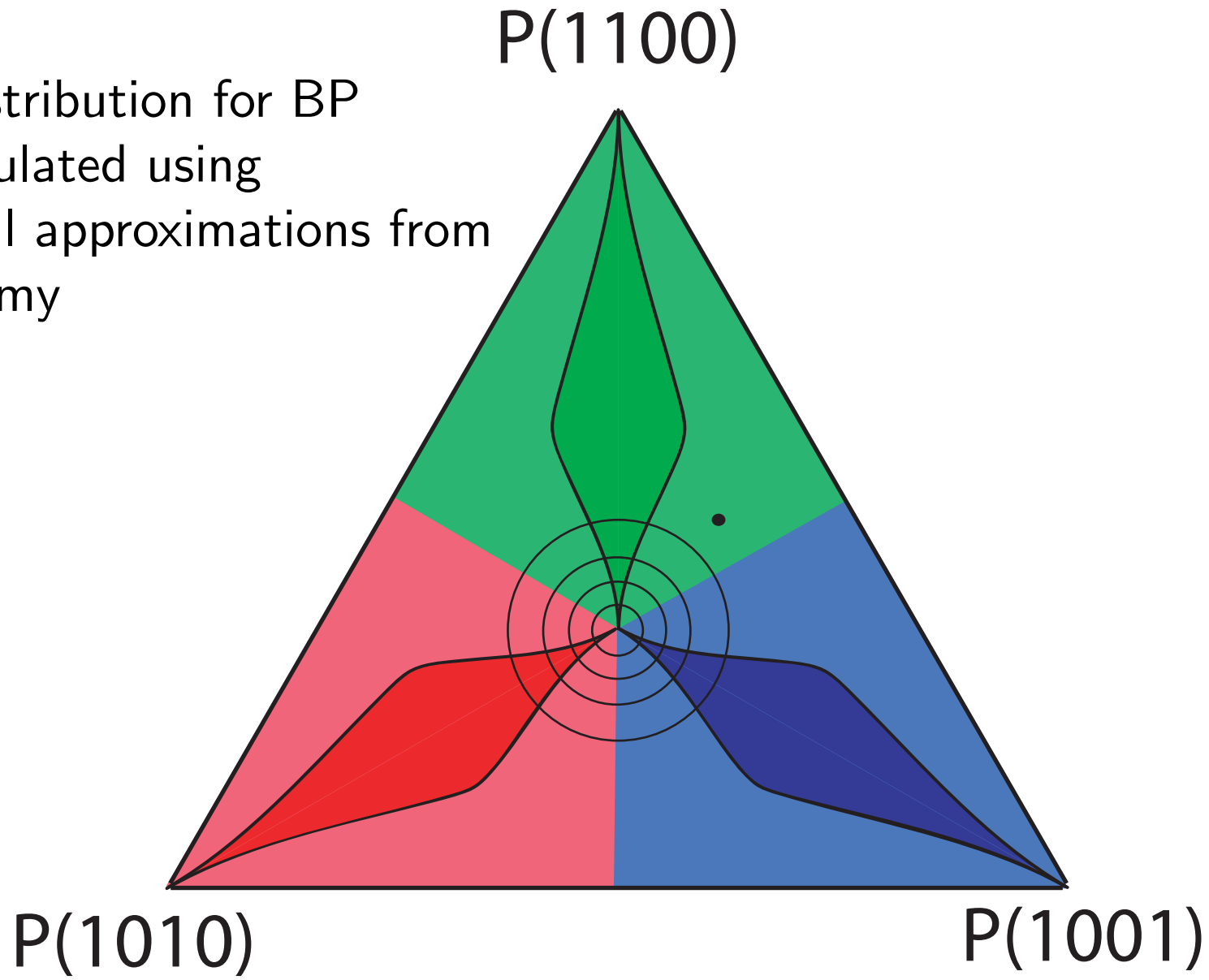
- Susko agrees with curvature arguments of Efron et al. (1996) and Shimodaira (2002), **but** points out that they ignore the **sharp point** in parameter space around the polytomy.
- He correct bootstrap proportions: $1 - aBP$ accurately estimates the P-value.
- The method uses the multivariate normal distributions the based on calculations about the curvature of the *likelihood* surface.
- You need to perform a different correction when you know the candidate tree *a priori* versus when you are putting BP on the ML tree.
- BP may **not** be conservative when you correct for selection bias.


aBP for each BP (5 model conditions)
aBP
BP
0
20
40
60
80
100


aBP with selection bias correction for each BP (5 model conditions)
aBP
BP
0
20
40
60
80
100

aBP in Pattern Frequency Space

Null distribution for BP is calculated using Normal approximations from polytomy



Summary - Part 1

- $\delta(T_1, T_2 \mid X) = 2\left[\ln L(T_1 \mid X) - \ln L(T_2 \mid X)\right]$ 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.

Summary - Part 2

A (very) wide variety of tests differ by:

- Null hypotheses:
  - Expected scores are the same $\rightarrow$ boundary tests. **Non-parametric tests**
  - A tree consistent with the null is correct $\rightarrow$ tests that use the full info of the model. **Parametric tests**

- How to use variance information:
  - Rely on "raw" bootstrap variability,
  - Invoke assumptions of normality of scores,
  - Use χ^2 variants.

- Whether or not the trees must be specified *a priori* – KH Test requires the trees to be specified *a priori*.

Summary - Part 3

	Parametric	Nonparametric
P-value from δ	aLRT, aBayes, parametric bootstrapping	KH, SH
P-value from BP	aBP(semi)	BP, aBP(semi), AU, EHH

When you use a parametric test, you will usually gain power. But non-parametric tests are more robust to model violation.

Significantly different genealogy $\neq$ different phylogeny

- True "gene tree" can differ from true "species tree" for several biological reasons:
  - deep coalescence,
  - gene duplication/loss (you may be comparing paralogs),
  - lateral gene transfer.

We often don't want to test tree topologies

- If we are conducting a "comparative method" we have to consider phylogenetic history,
- ideally we would integrate out the uncertainty in the phylogeny,
- this entails averaging over trees, but not averaging P-values (or point estimates) over trees.

References

Alfaro, M. E., Zoller, S., and Lutzoni, F. (2003). Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, 20(2):255–266.

Anisimova, M. and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539–552.

Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C., and Gascuel, O. (2011). Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*.

Buckley, T. R. (2002). Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Systematic Biology*, 51(3):509–523.

Cummings, M. P., Handley, S. A., Myers, D. S., Reed, D. L., Rokas, A., and Winka, K. (2003). Comparing bootstrap and posterior probability values in the four-taxon case. *Systematic Biology*, 52(4):477–487.

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., and Douzery, E. J. P. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20(2):248–254.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Science, U. S. A.*, 93:13429–13434.

Goldman, N., Anderson, J. P., and Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49:652–670.

Hillis, D. M. and Bull, J. J. (1993). An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology*, 42(2):182–192.

Huelsenbeck, J., Hillis, D., and Nielsen, R. (1996). A Likelihood-Ratio Test of Monophyly. *Systematic Biology*, 45(4):546.

Huelsenbeck, J. P. and Rannala, B. (2004). Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53:904–913.

Kim, J. (2000). Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. *Molecular Phylogenetics and Evolution*, 17(1):58–75.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31:151–160.

Newton, M. A. (1996). Bootstrapping phylogenies: Large deviations and dispersion effects. *Biometrika*, 83(1):315–328.

Ota, R., Waddell, P. J., Hasegawa, M., Shimodaira, H., and Kishino, H. (2000). Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution*, 17(5):798–803.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, 57(5):758–771.

Strimmer, K. and von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 94(13):6815–6819.

Susko, E. (2010). First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. *Molecular Biology and Evolution*, 27(7):1621–1629.

Susko, E. (2014). Tests for two trees using likelihood methods. *Molecular Biology and Evolution*.

Suzuki, Y., Glazko, G. V., and Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *PNAS*, 99:16138–16143.

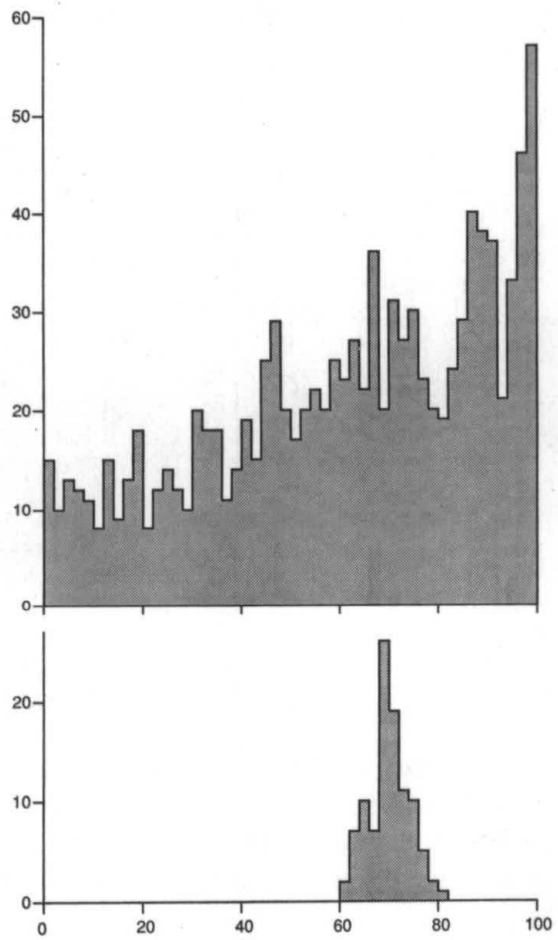
Svennblad, B., Erixson, P., Oxelman, B., and Britton, T. (2006). Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics. *Systematic Biology*, 55(1):116–121.

Wilcox, T. P., Zwickl, D. J., Heath, T., and Hillis, D. M. (2002). Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Molecular Phylogenetics and Evolution*, 25:361–371.

Bootstrapping as a noisy measure of repeatability

bootstrap values from many simulations

repeated simulation



% recovering tree

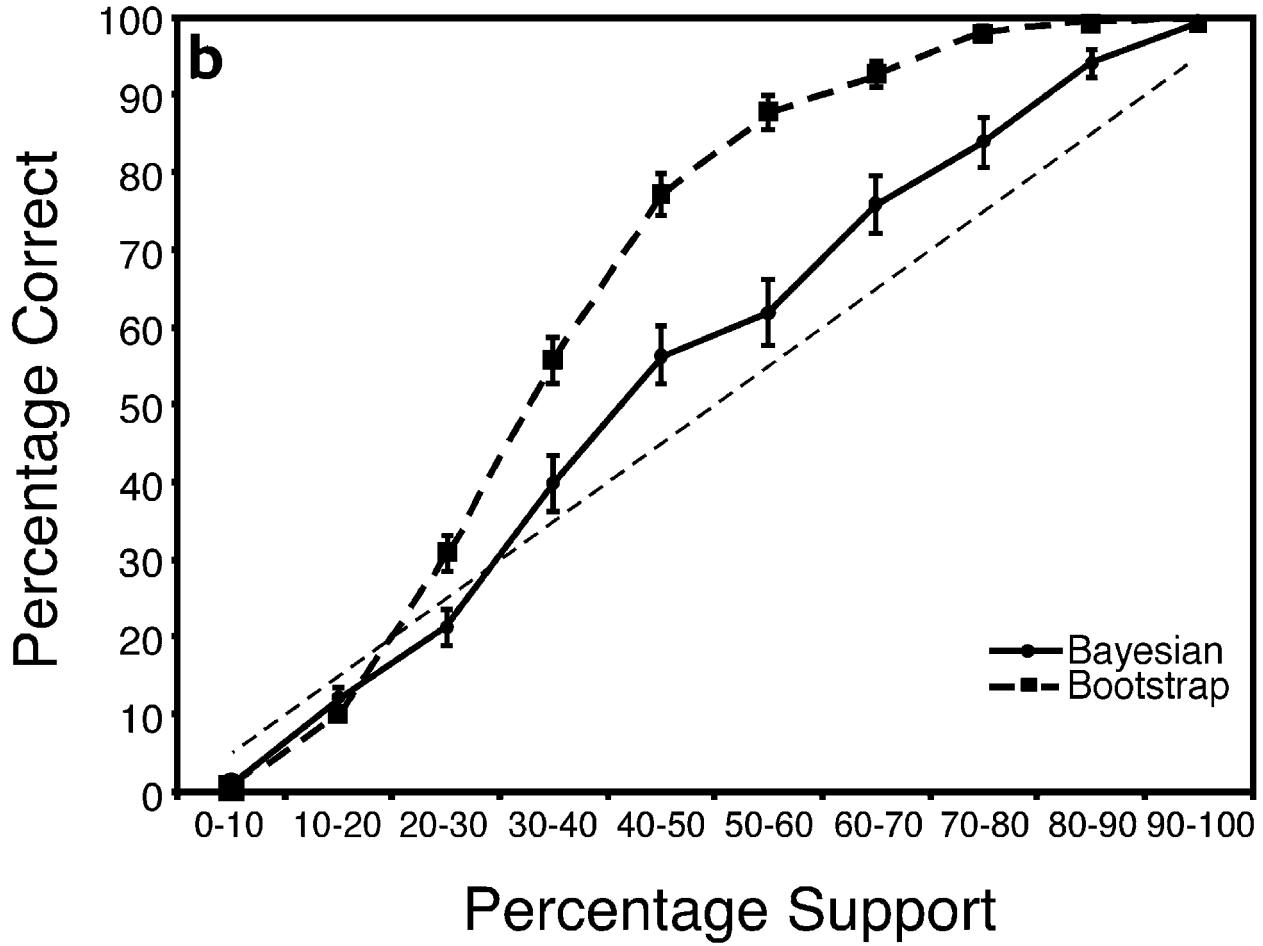
Simulation study of Hillis and Bull (1993)

Bootstrap Proportion $\neq$ Posterior Probability

Several studies have compared the non-parametric bootstrap proportion of clade from an ML analysis of a data set to the posterior probabilities when the same data is analyzed under the same model (Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003).

Note: **Not** all of these have implied that the measures **should** be the same, but some authors have (usually citing Efron et al., 1996).

Bootstrap Proportion $\neq$ Posterior Probability in general



from Wilcox et al. (2002)

Note: Huelsenbeck and Rannala (2004) showed that the Bayesian posterior pro
are right on the equality line, if you simulate from the prior.

Newton (1996) showed that, when you look at the median, the BP may not be biased downward

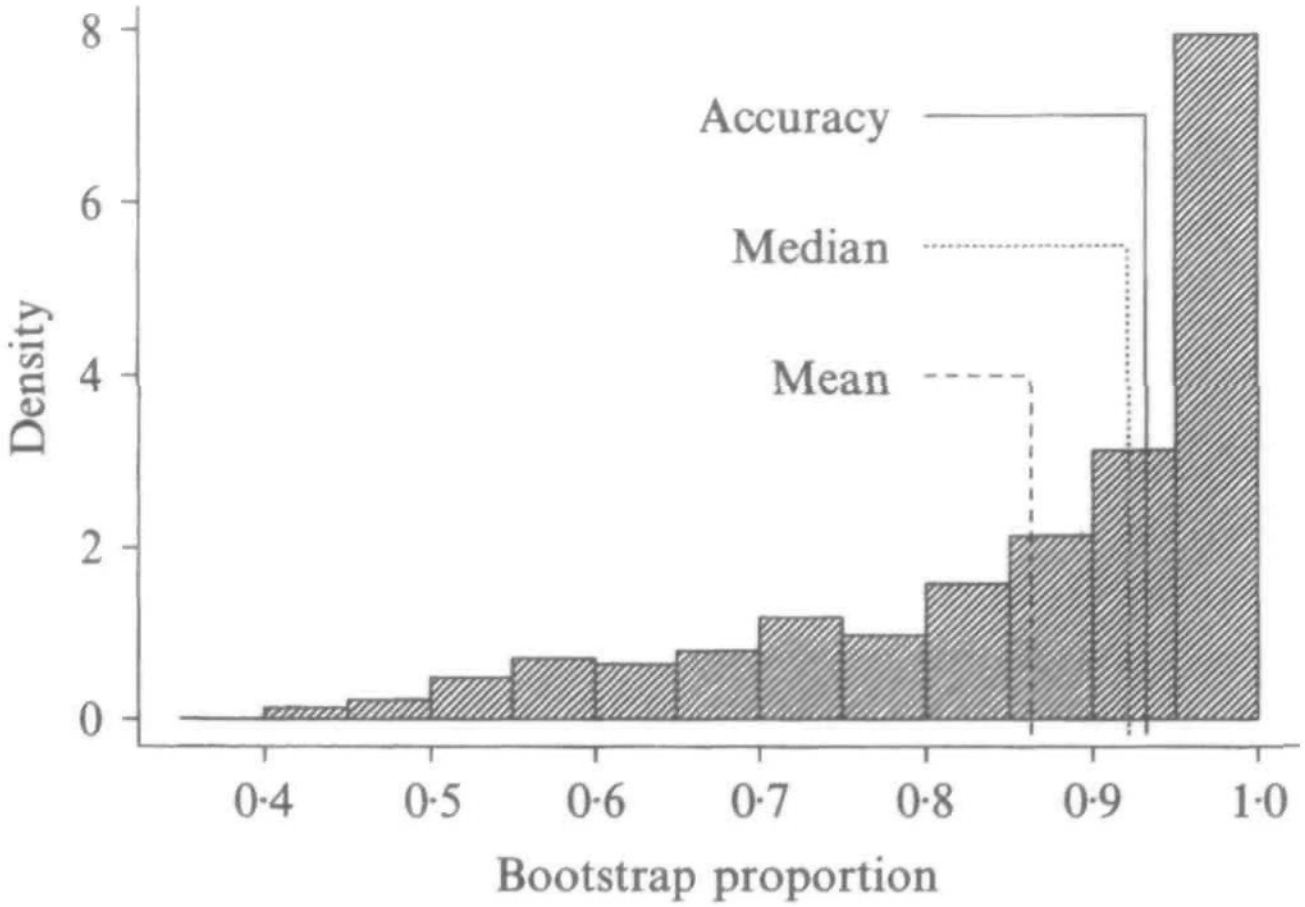
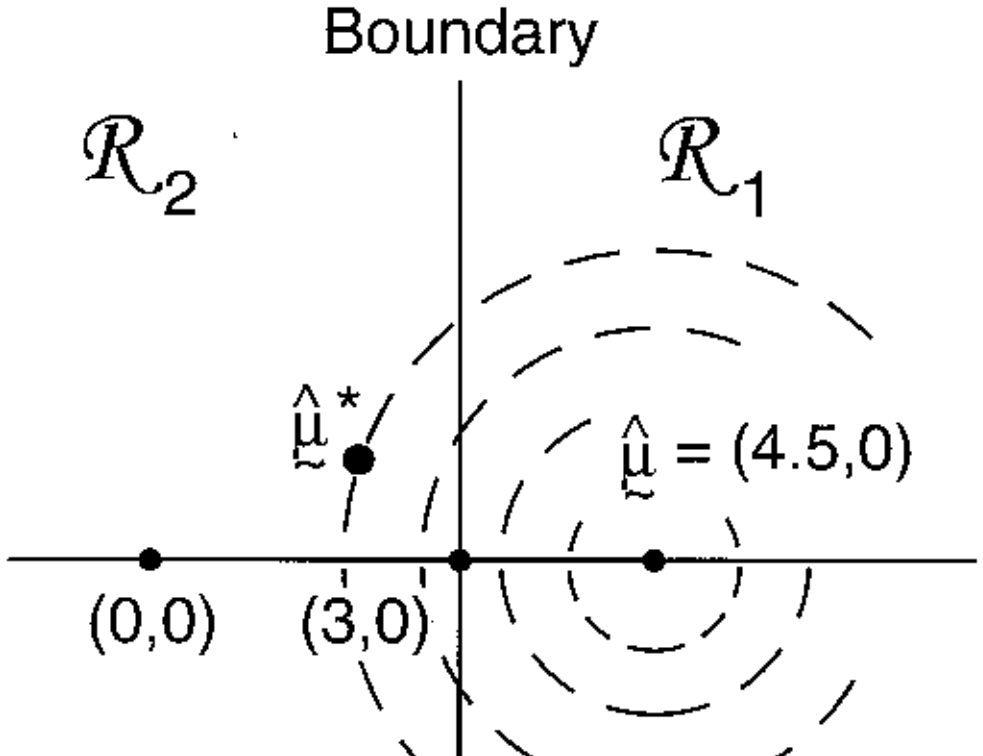


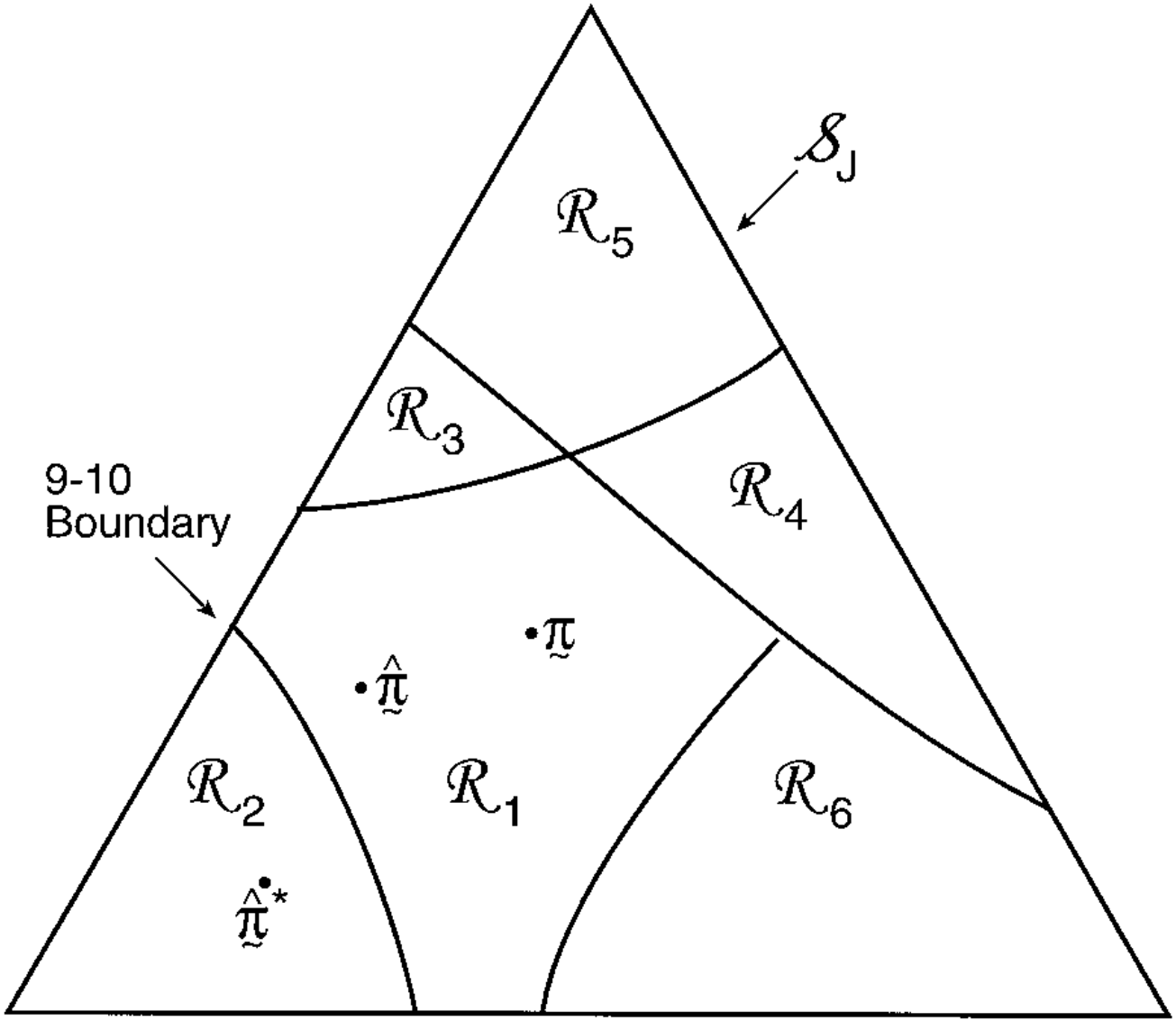
Figure 4 from Newton (1996)

What did Efron et al. (1996) say?

We can use a Bayesian model to show that $\tilde{\alpha}$ is a reasonable assessment of the probability that $\mathscr{R}_1$ contains μ. Suppose we believe $a\ priori$ that μ could lie anywhere in the plane with equal probability. Then having observed $\hat{\mu}$, the $a\ posteriori$ distribution of μ given $\hat{\mu}$ is $N_2(\hat{\mu}, I)$ exactly the same as the bootstrap distribution of $\hat{\mu}^*$. In other words, $\tilde{\alpha}$ is the $a\ posteriori$ probability of the event $\mu \in \mathscr{R}_1$, if we begin with an "uninformative" prior density for μ.



Efron et al. (1996) view of tree space



What did Efron et al. (1996) say (and mean)?

- the "uninformative" prior density is a uniform prior over all of pattern frequency space

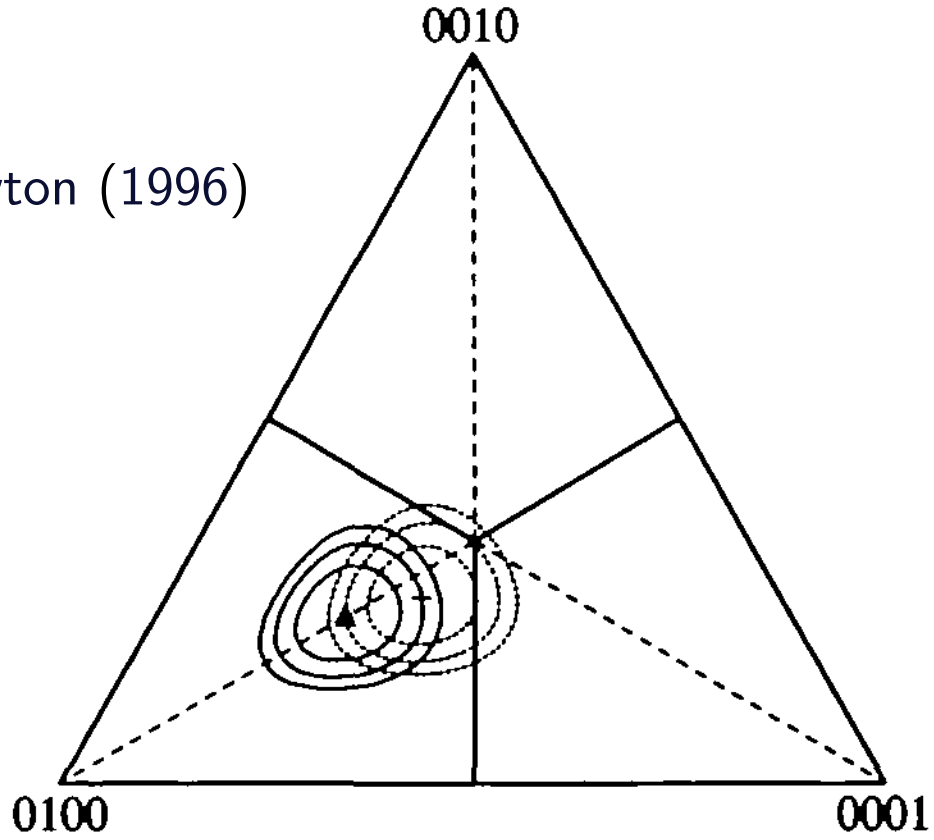
- this is *not* equivalent to a prior that would be expected to yield a phylogeny (it is actually identical to the prior you would get if you assumed that all pairwise distances between taxa were ∞),

- Efron et al. (1996) were *not* predicting that the bootstrap proportions should be identical to those from a Bayesian phylogenetic analysis with real phylogenetic priors.

- Svennblad et al. (2006) have a nice paper on this subject.

Newton (1996) provides an intuition for why the mean BP may be lower than repeatability

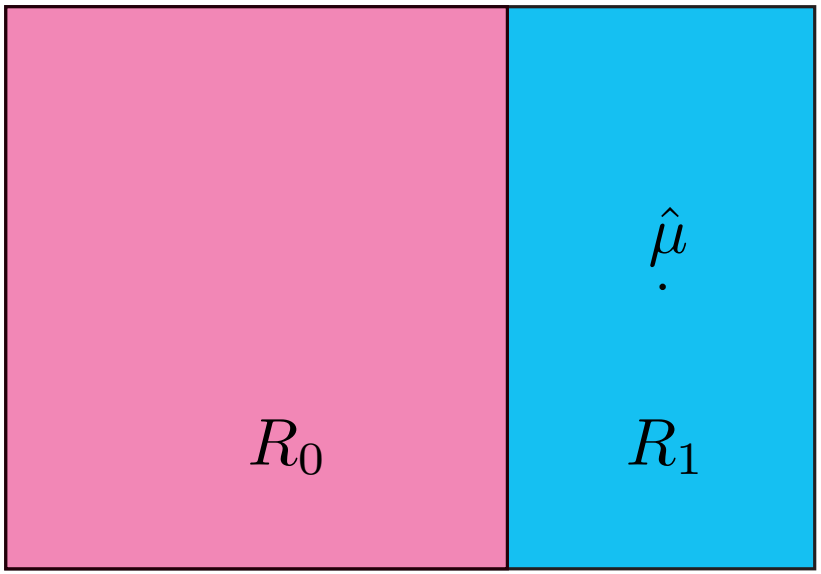
Figure 3 from Newton (1996)



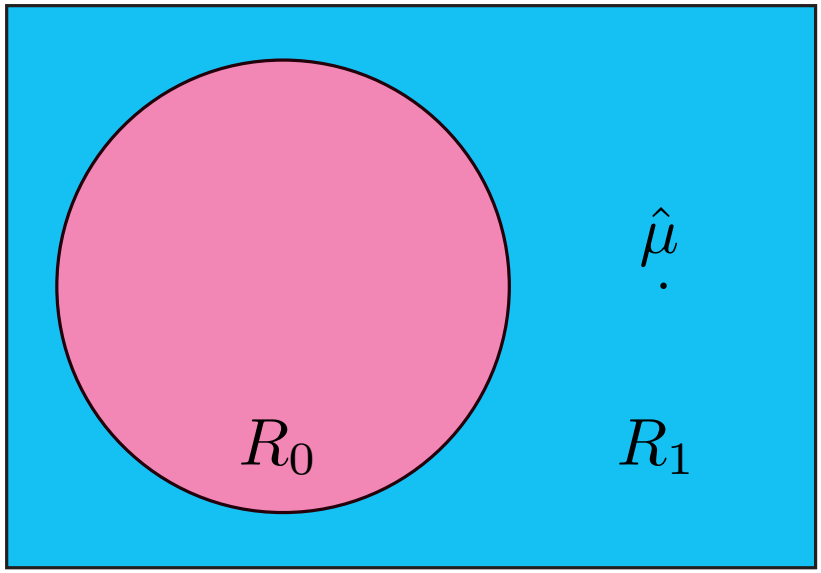
Darker ovals indicate probability contours for datasets given the truth (note that repeatability $\approx 100\%$)

Lighter ovals show probability contours for bootstrapping for one dataset. Many real datasets will have BP much $< 100\%$

Case 1

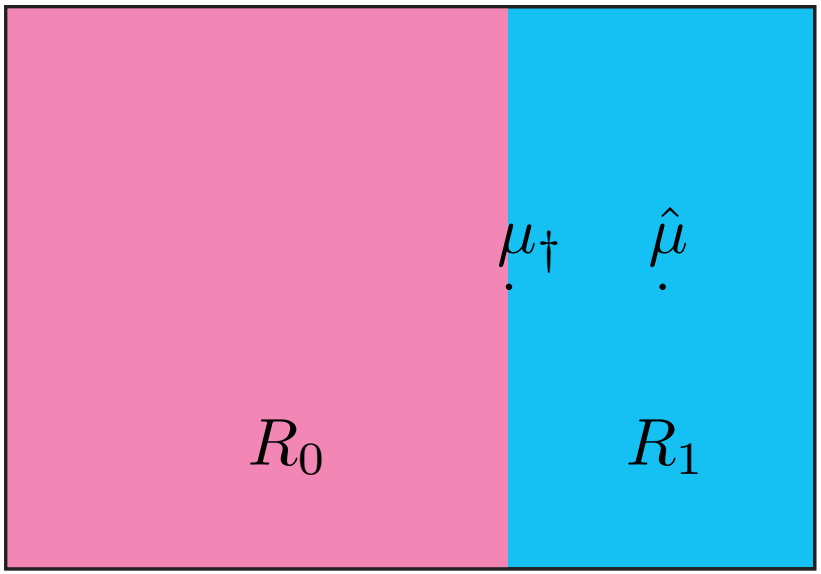


Case 2

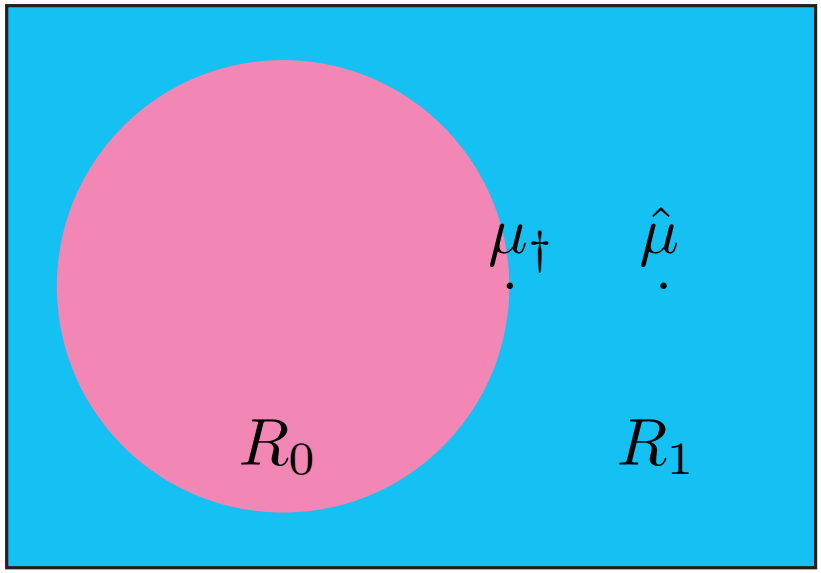


$\hat{\mu}$ is the best point calculated from the data

Case 1

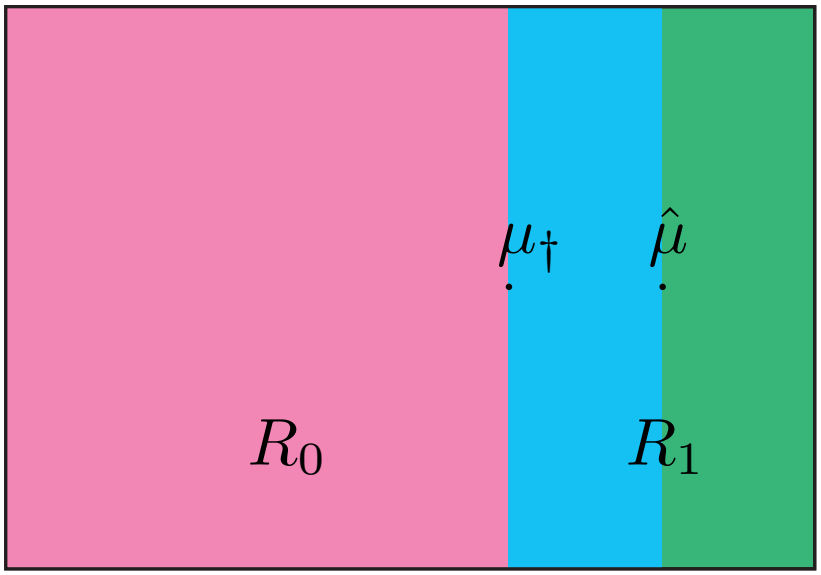


Case 2

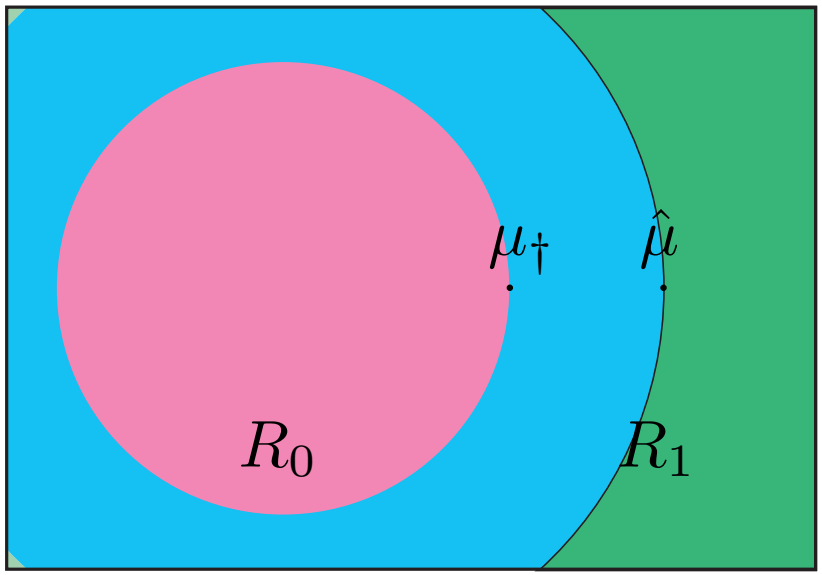


$\hat{\mu}$ is the best point calculated from the data
$\mu_\dagger$ is least-favorable condition (LFC) point in R_0

Case 1

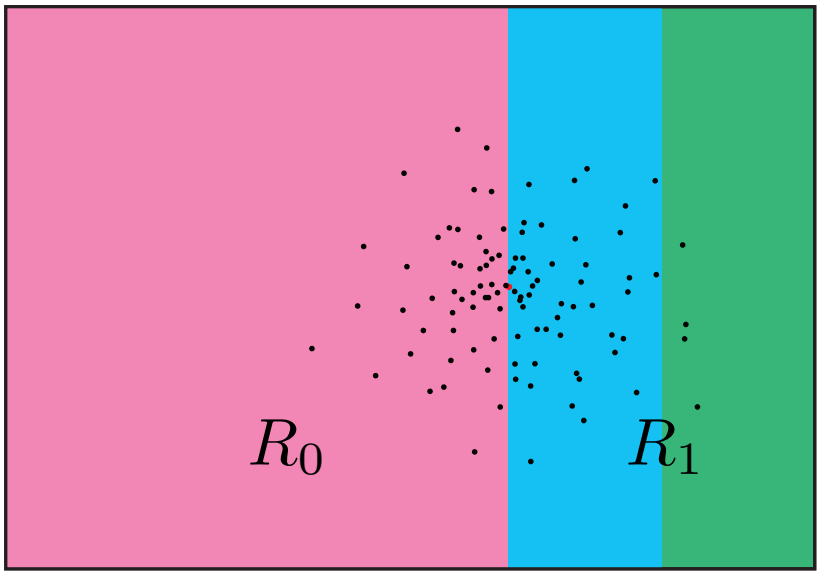


Case 2

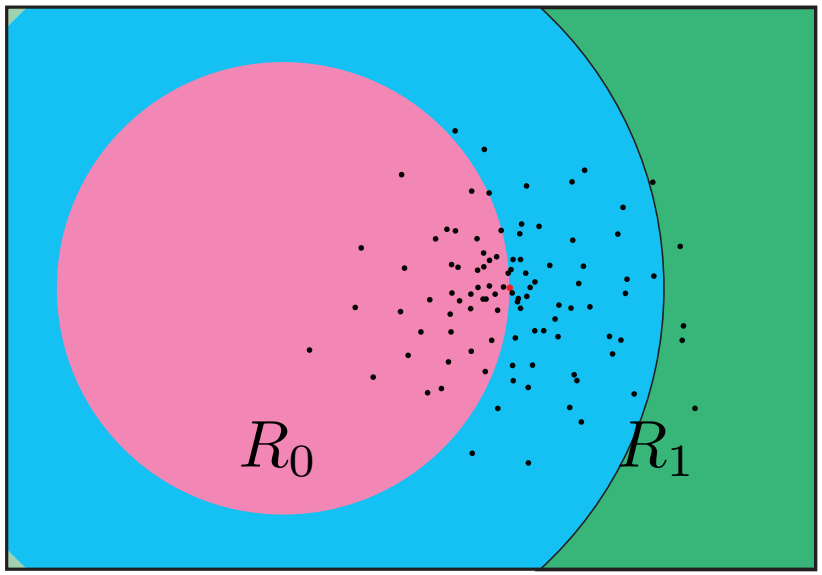


$\hat{\mu}$ is the best point calculated from the data
$\mu_\dagger$ is least-favorable condition (LFC) point in R_0
green areas are the tails - they correspond to values of the test statistic more extreme than $\hat{\mu}$ (relative to that $\mu \in R_0$

Case 1

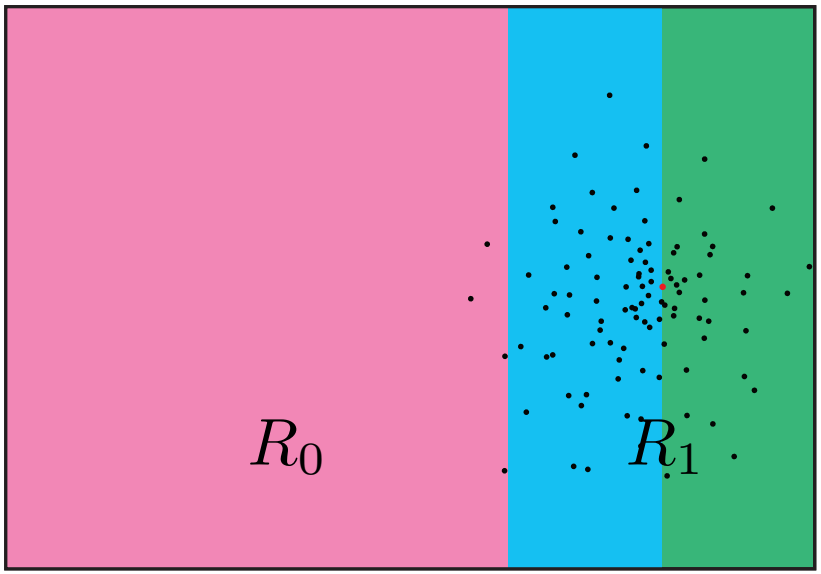


Case 2

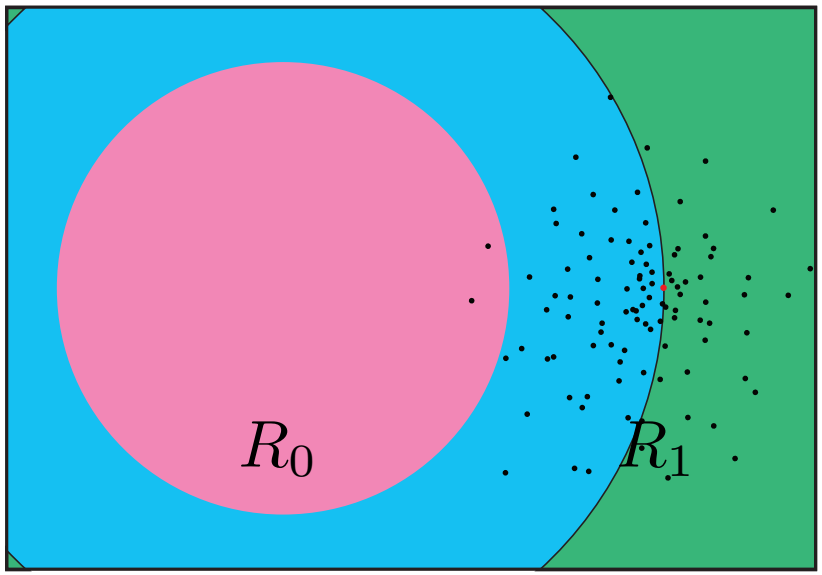


$\hat{\mu}$ is the best point calculated from the data
$\mu_\dagger$ is least-favorable condition (LFC) point in R_0
Case 1 P-value $<$ the P-value in Case 2

Case 1

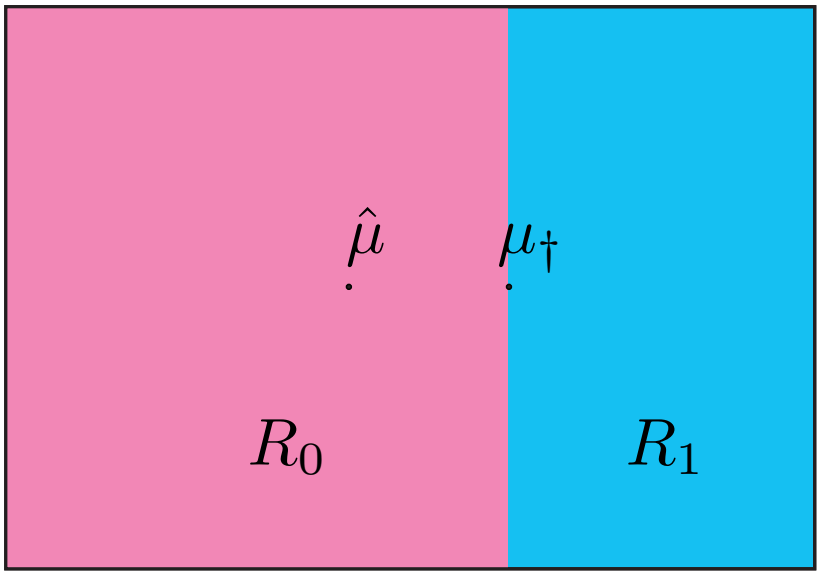


Case 2

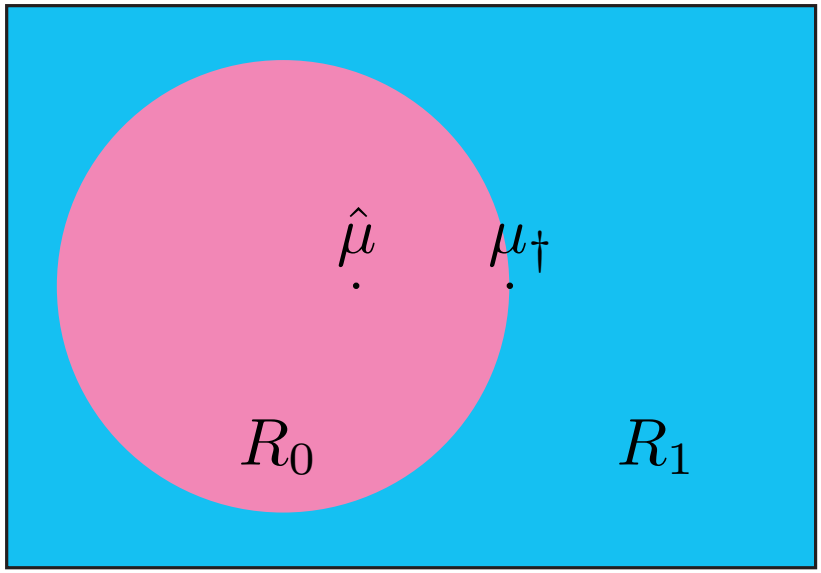


In case 1 - the bootstrap proportion is a good estimate of the P-value
In case 2 - the bootstrap proportion underestimates the P-value

Case 3

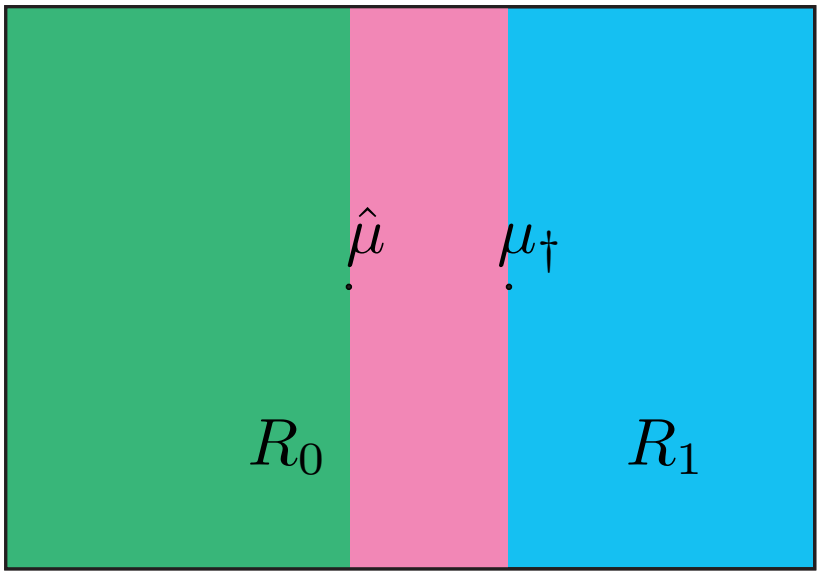


Case 4

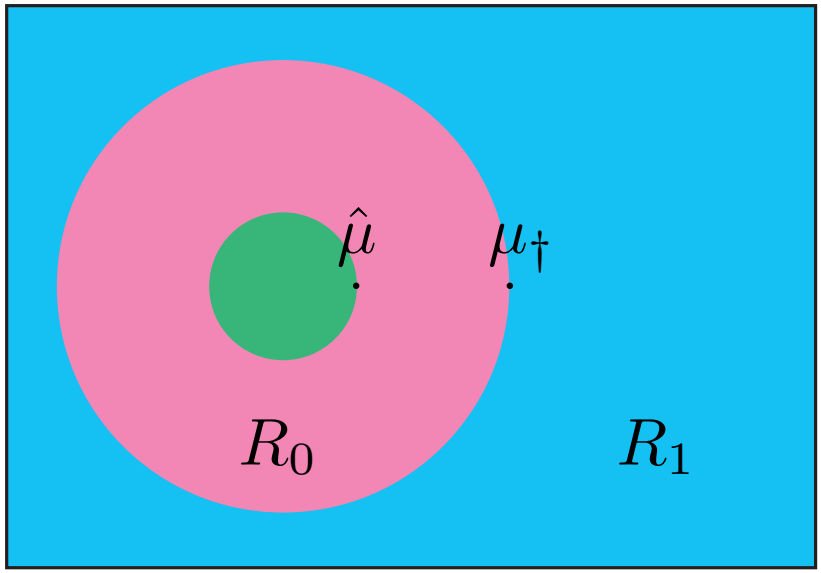


$\hat{\mu}$ is the best point calculated from the data
$\mu_\dagger$ is least-favorable condition (LFC) point in R_1

Case 3

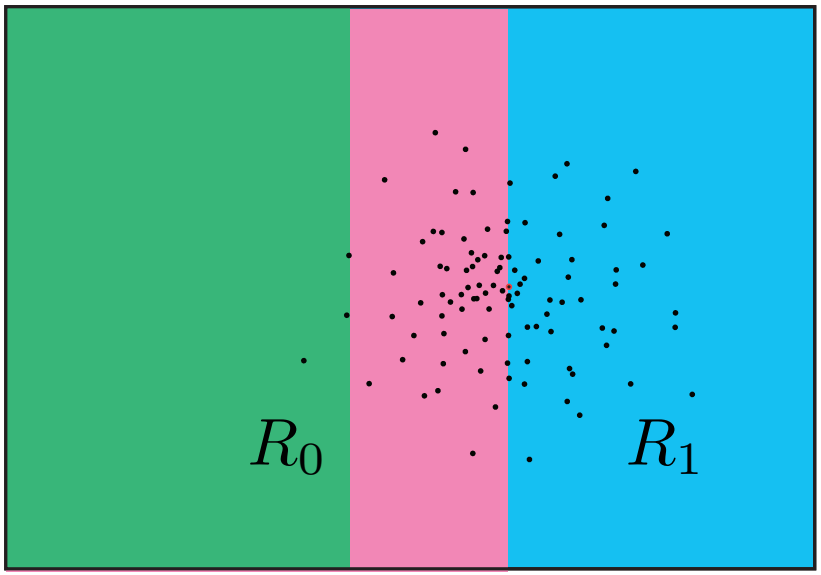


Case 4

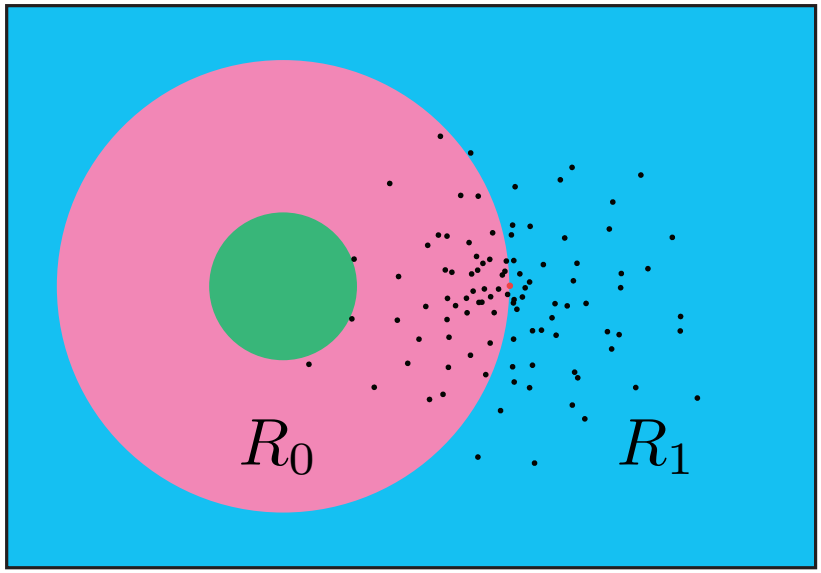


$\hat{\mu}$ is the best point calculated from the data
$\mu_\dagger$ is least-favorable condition (LFC) point in R_0
green areas are the tails - they correspond to values of the test statistic more extreme than $\hat{\mu}$ (relative to that $\mu \in R_1$)

Case 3

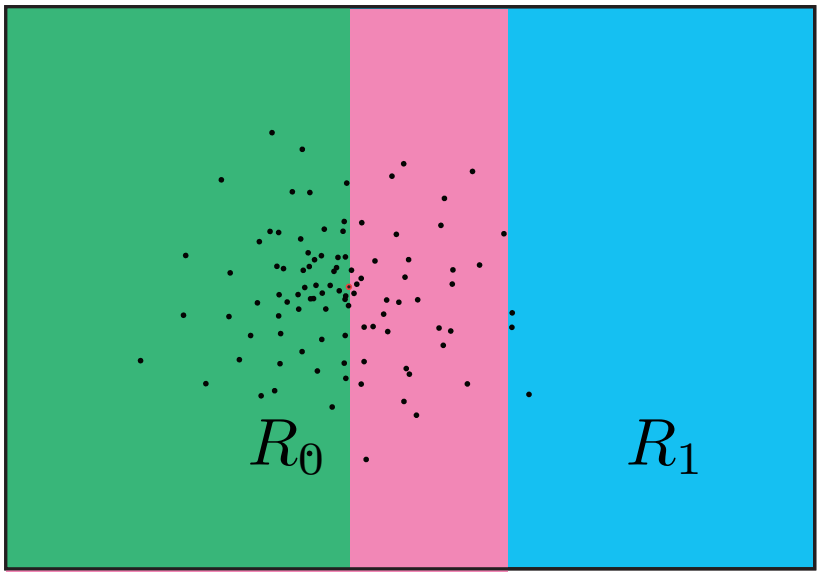


Case 4

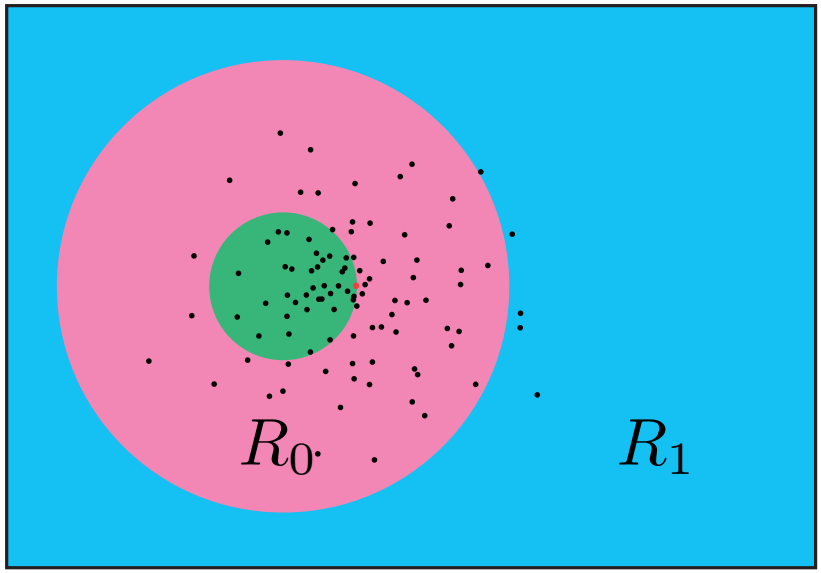


Case 3 P-value $>$ the P-value in Case 4

Case 3



Case 4

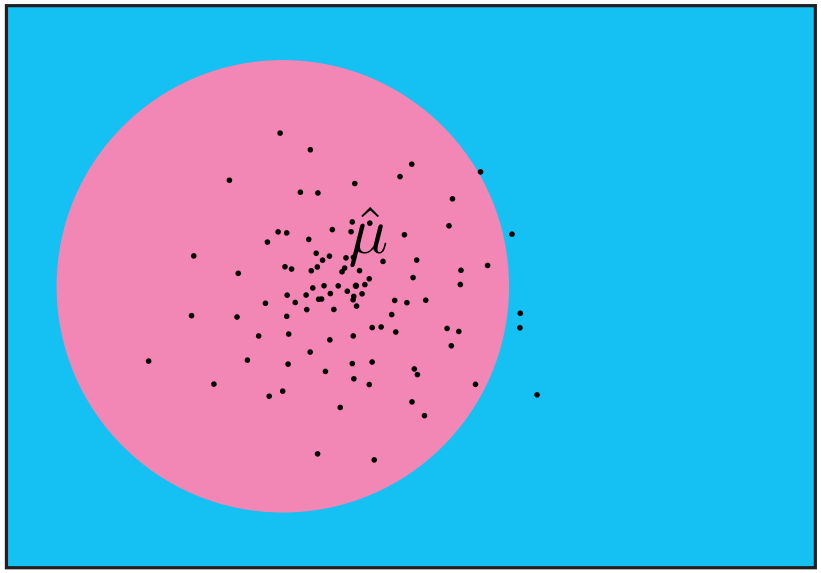


In case 3 - the bootstrap proportion is a good estimate of the P-value
In case 4 - the bootstrap proportion overestimates the P-value

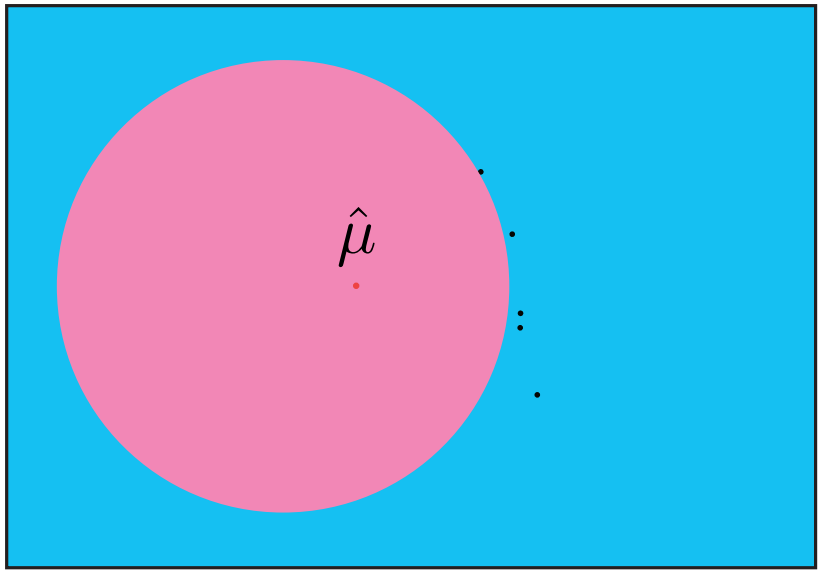
Efron et al. (1996) pointed out these issues of curvature of the boundaries between tree hypotheses.

We cannot see the boundaries in tree space, so it is hard to know how to correct for the biases so that we can use bootstrapping procedures as a means of getting a P-value for a clade – the probability that we would see this much support (or stronger support) for a clade if it were *not* present in the true tree.

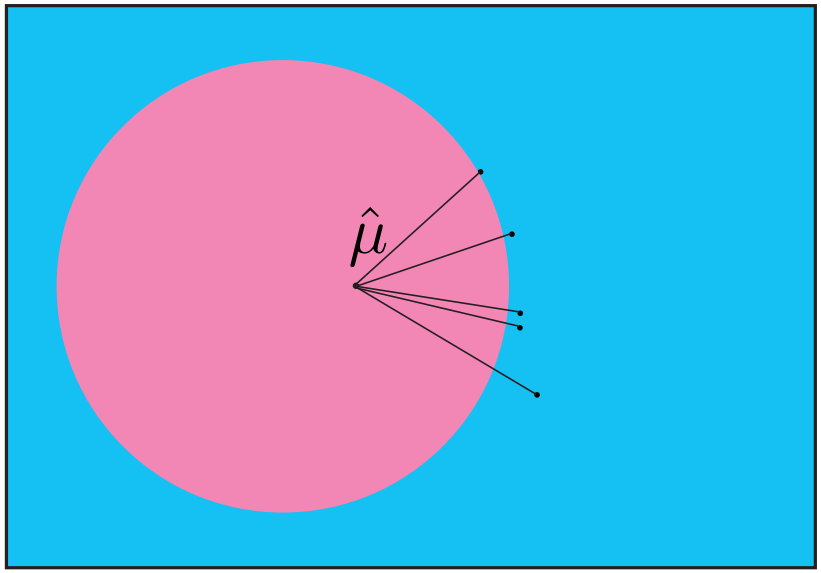
Initial bootstrap



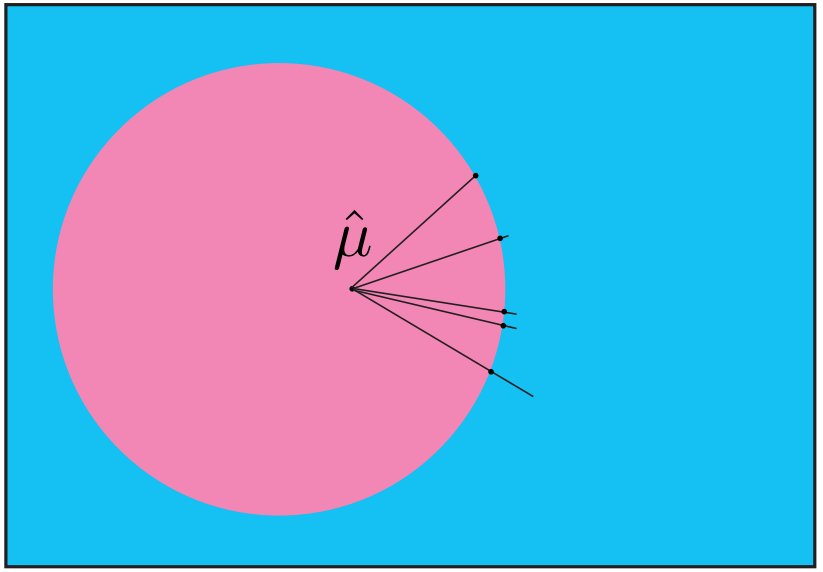
Find replicates that return a tree without the clade



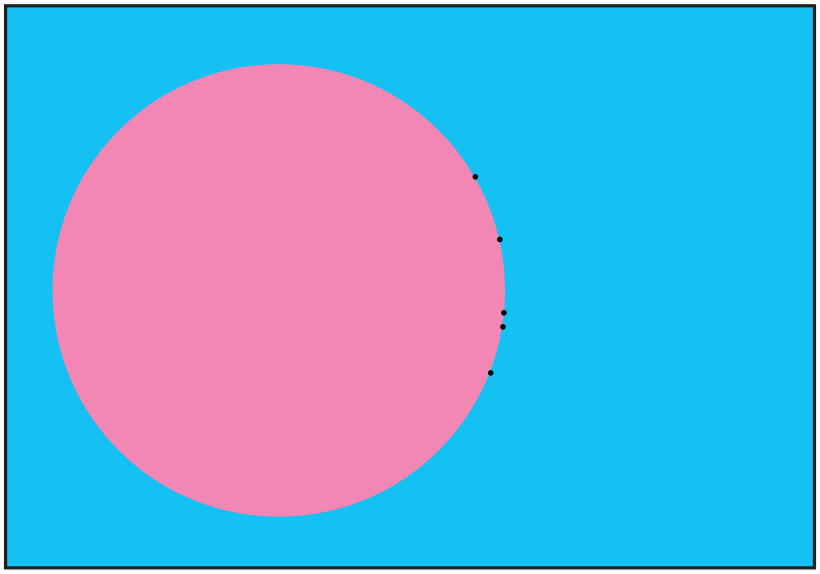
Find replicates that return a tree without the clade



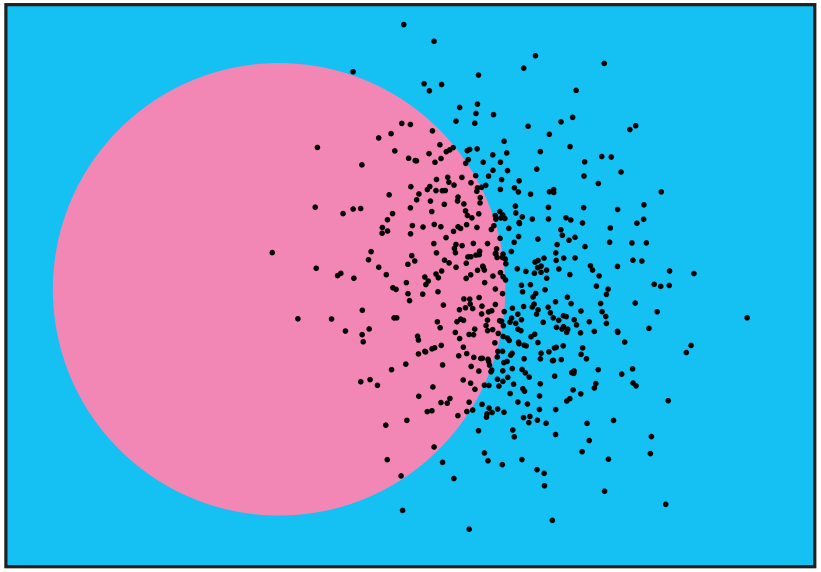
Find boundary points between regions



Find boundary points between regions



Bootstrap from these boundary points to check curvature of the boundary



The corrected bootstrap procedure of Efron et al. (1996) requires a very large number of bootstrap replicates because you need very accurate estimates of the curvature in order to apply the correction. Shimodaira (2002) expanded on this work:

- d is the distance from the point that corresponds to the data and the closest point on the boundary between another tree
- $\Phi(\cdot)$ denotes the cumulative density function of the standard Normal(0,1) distribution.
- c denotes the curvature of the boundary
- the P-value for the KH test is given by $KH = \Phi(d)$

- Shimodaira argues (from an early Efron paper) that the appropriate P-value for tree selection is:

$$AU = 1 - \Phi(d - c)$$

- In "standard" non parametric bootstrapping proportions are:

$$BP = 1 - \Phi(d + c)$$

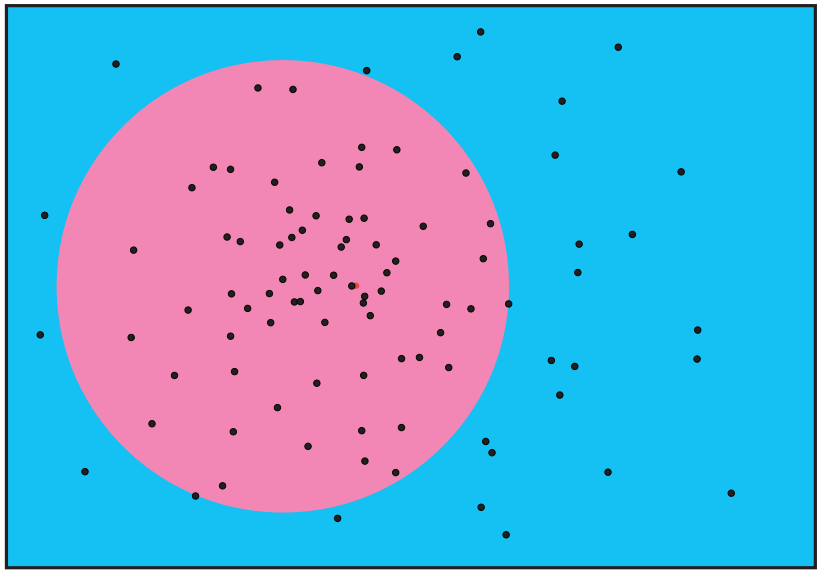
  Note the incorrect sign with respect to the curvature term causes BP (and recall how on the curved boundary examples, the curvature caused the P-value to change in one direction and the BP to go in the other).

How can we find c so that we can correct for it?

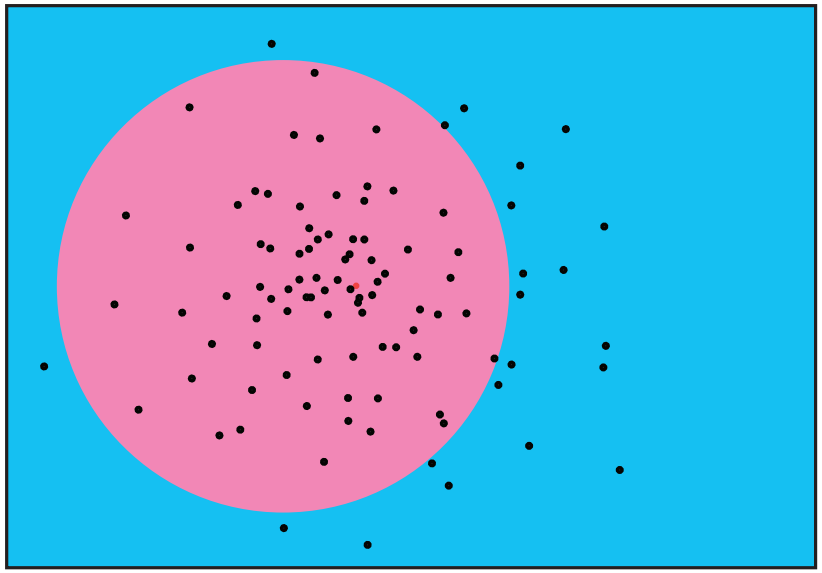
- N is the number of characters in the real data set
- N' is the number of characters in each bootstrap data set
- $r = \frac{N'}{N}$
- If you do a bootstrap in which $r \neq 1$, Shimodaira determined the expected effect on the bootstrap proportion as a function of d and c:

$$BP(r) = 1 - \Phi\left(d\sqrt{r} + \frac{c}{\sqrt{r}}\right)$$

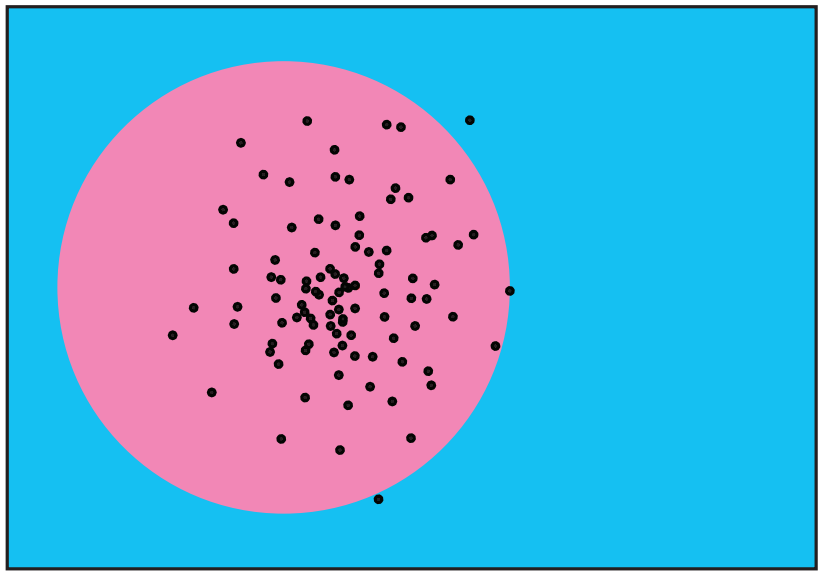
r = 0.5



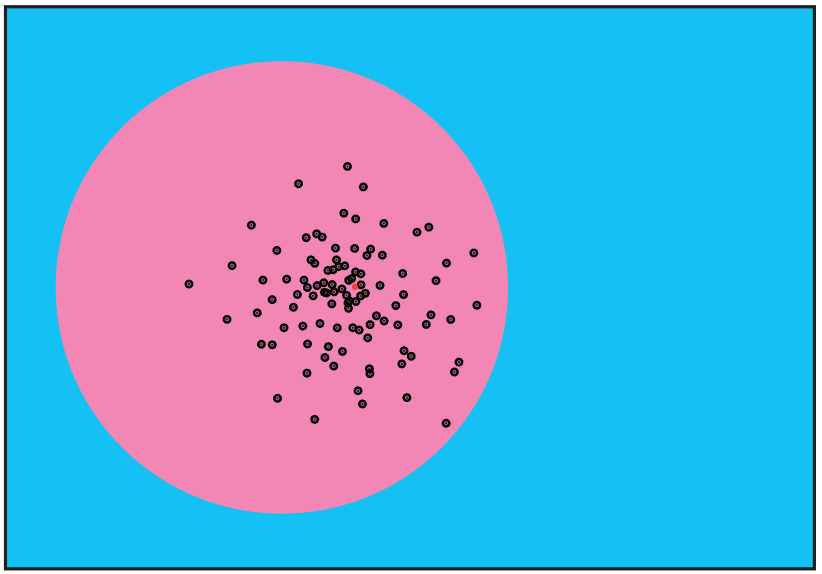
r = 0.8



r = 1.0



r = 1.2



AU Test

1. conduct a sweep of bootstraps with r varying (for instance $r = 0.5, r = 0.6, r = 0.7, \ldots r = 1.4$, to get a set of $BP(r)$ for a tree.
2. Use weighted least squares to estimate c and d form the set of $BP(r)$
3. Calculate

$$AU = 1 - \Phi(d - c)$$

This lets you calculate a P-value for any tree of interest, and then you can construct a confidence set of trees.