

Testing phylogenetic hypotheses

Woods Hole Workshop on Molecular Evolution, 2012

Mark T. Holder
University of Kansas

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

Reasons phylogenetic inference might be wrong

1. Our data might be wrong (“garbage in garbage out”)
2. *Systematic error* – Our inference method might not be sophisticated enough
3. 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 evening: **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,
3. Brief lab on parametric bootstrapping.
4. Continue the onslaught of testing methods
 - aLRT, aBayes,
 - 1 - BP,
 - AU and Efron et al. (1996) correction
 - aBP
5. Cartoon time! (warning: “cartoon time” will not actually be fun)

Some resources related to this talk

A [Mendeley Group](#) of papers related to topology testing on trees.

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

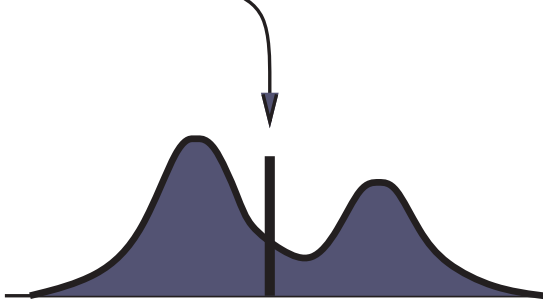
The source for all the documents for my talk is at:

<https://github.com/mtholder/TreeTopoTestingTalks>

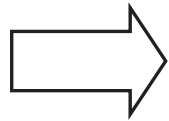
The bootstrap

(unknown) true value of

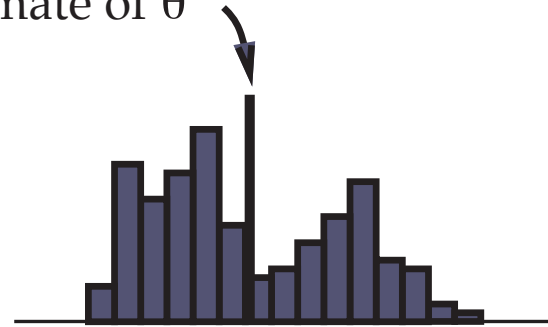
θ



(unknown) true distribution

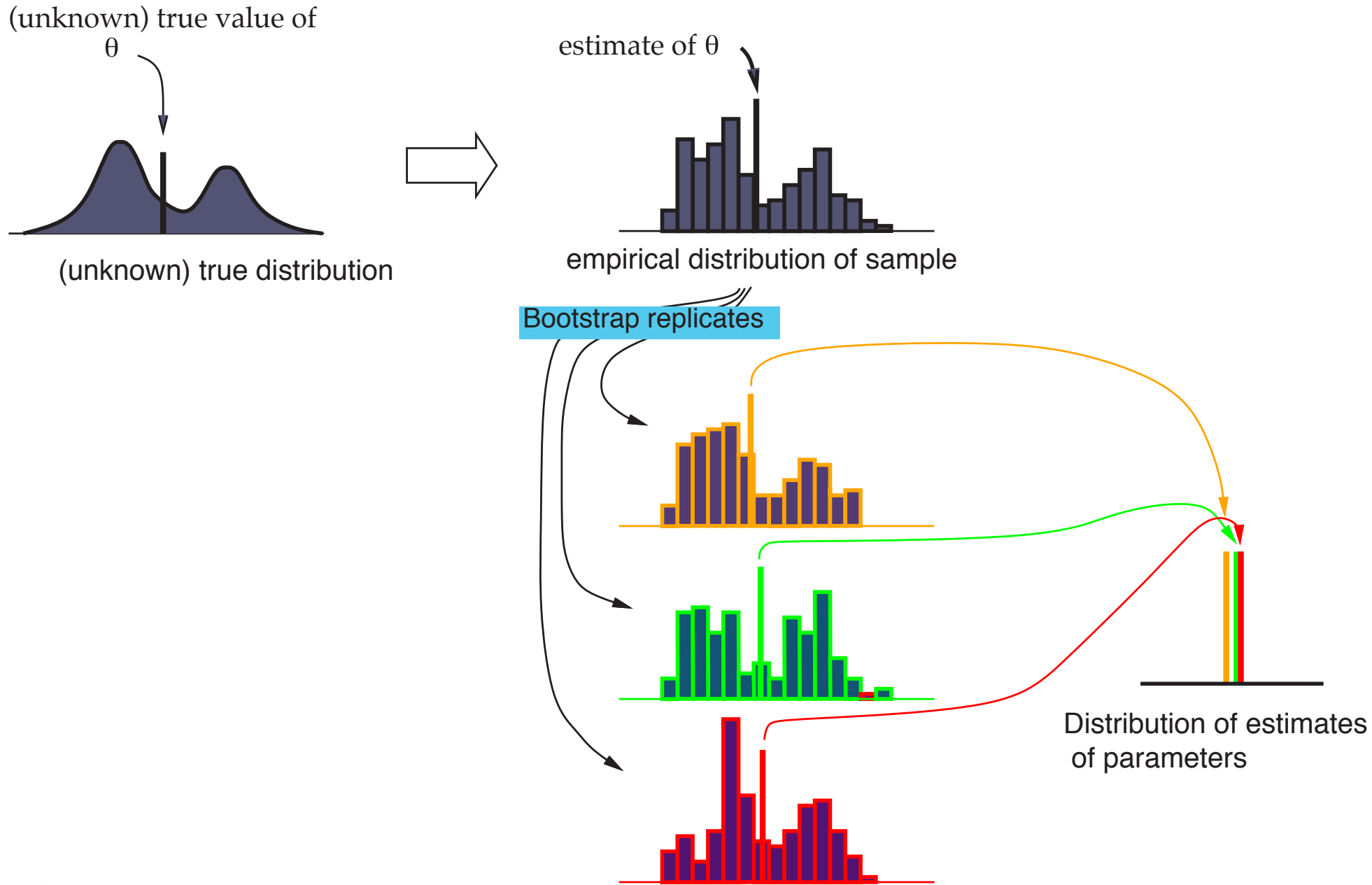


estimate of θ

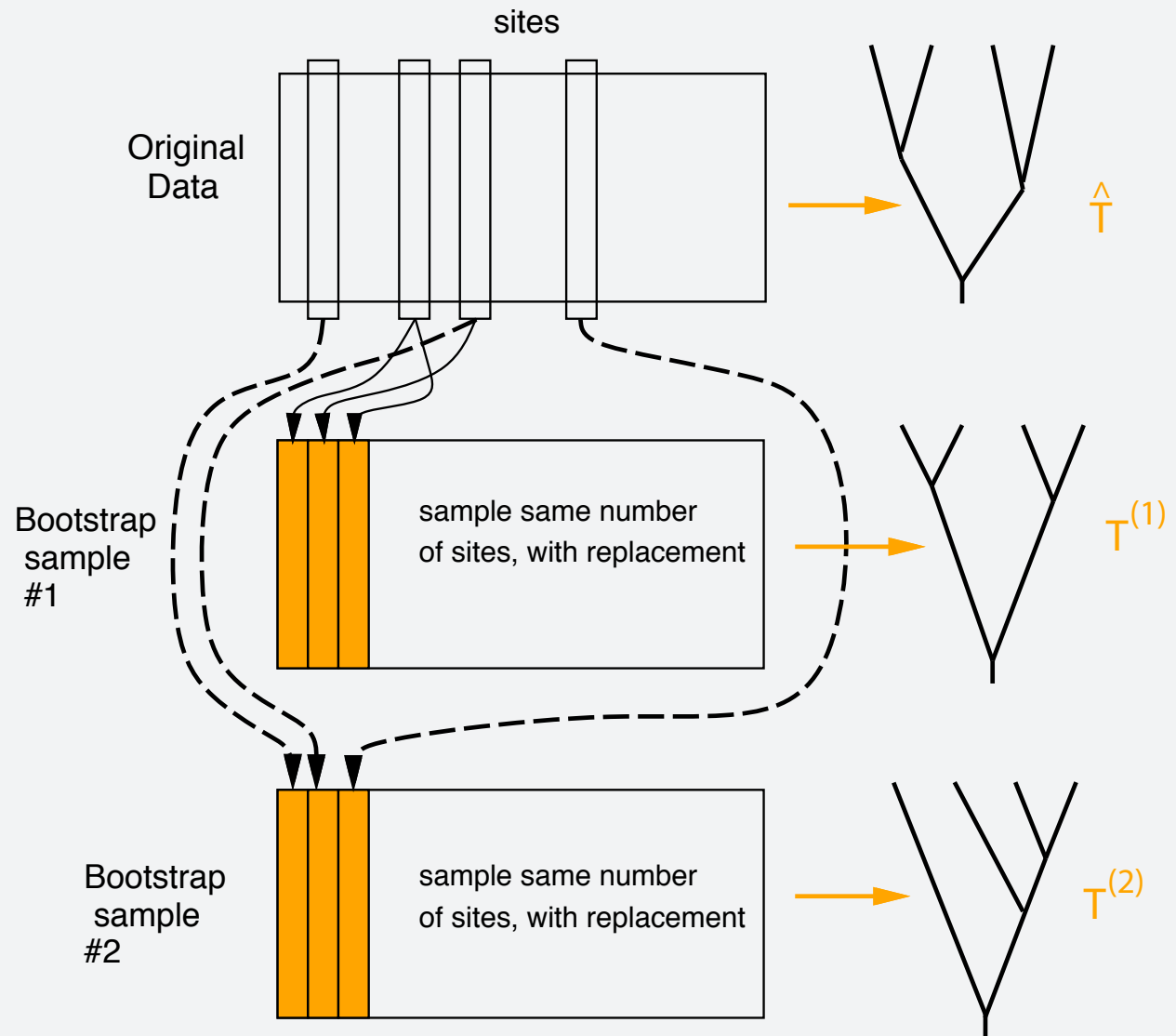


empirical distribution of sample

The bootstrap



The bootstrap for phylogenies

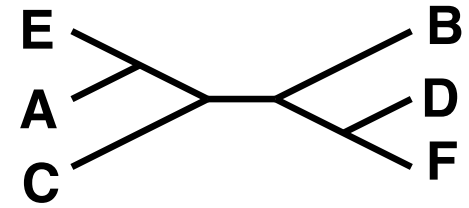
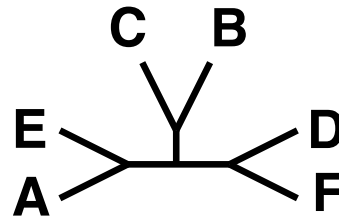
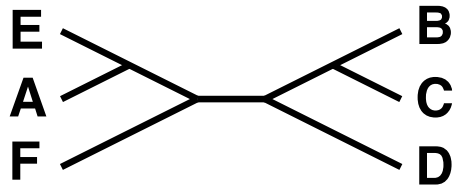
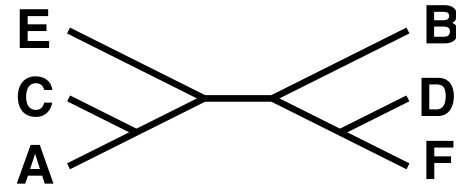
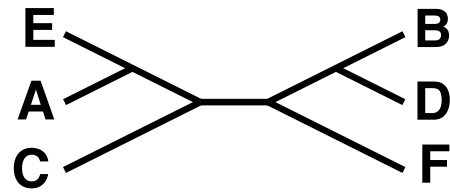


Slide from Joe Felsenstein

(and so on)

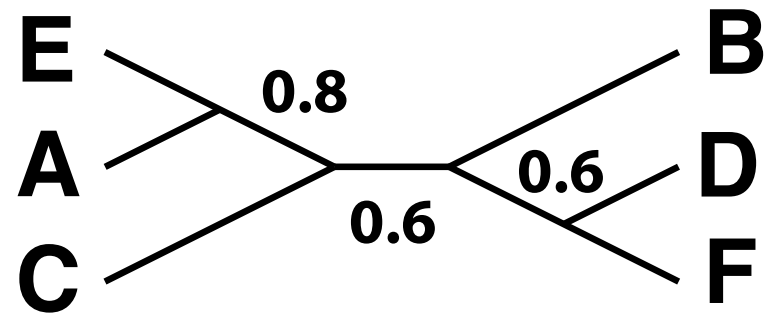
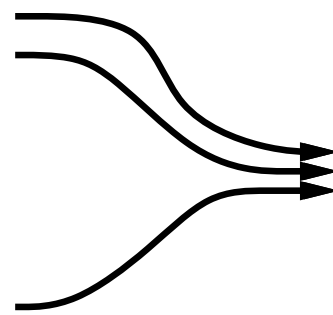
The majority-rule consensus tree

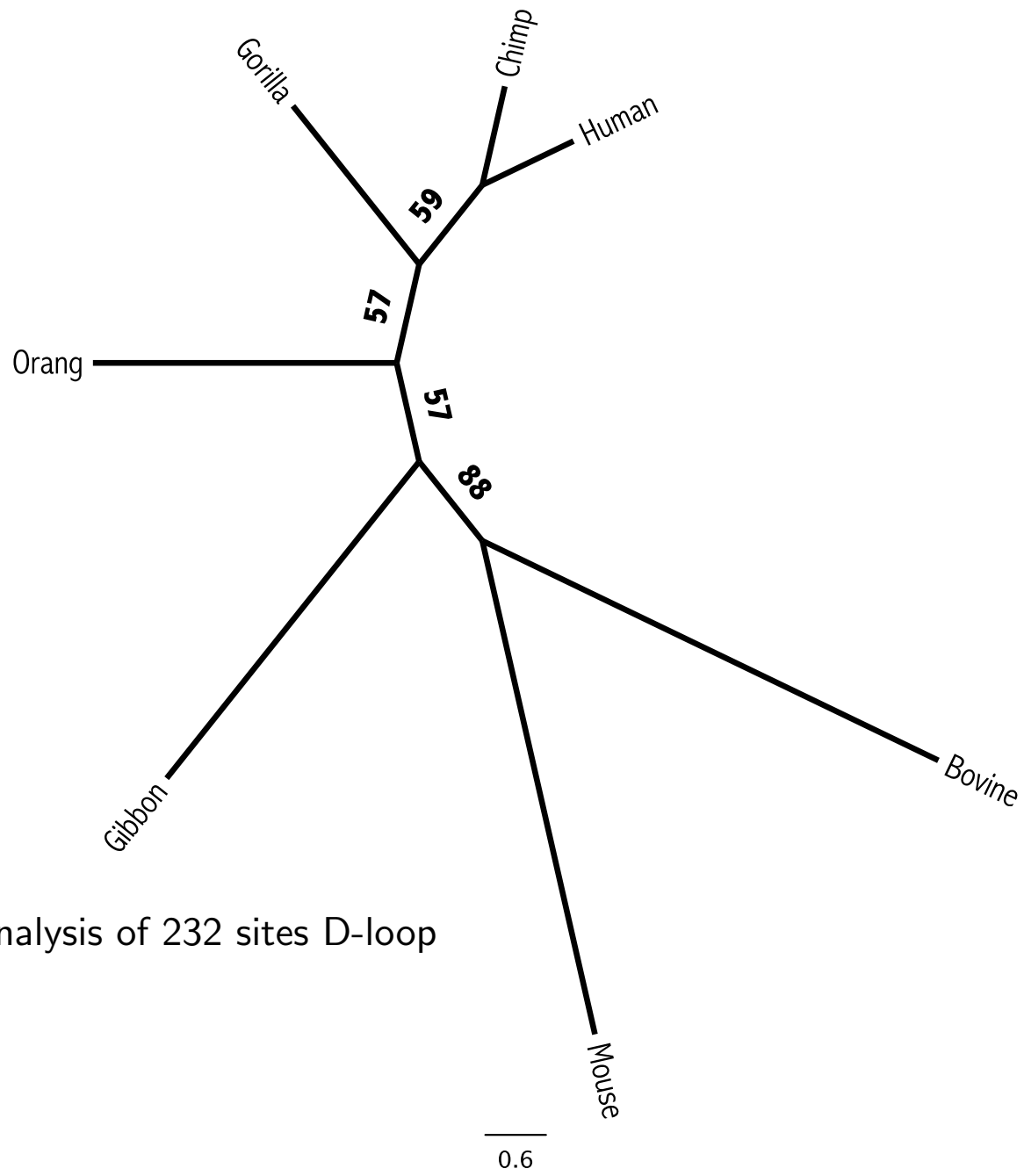
Trees:



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





From Hasegawa's analysis of 232 sites D-loop

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.

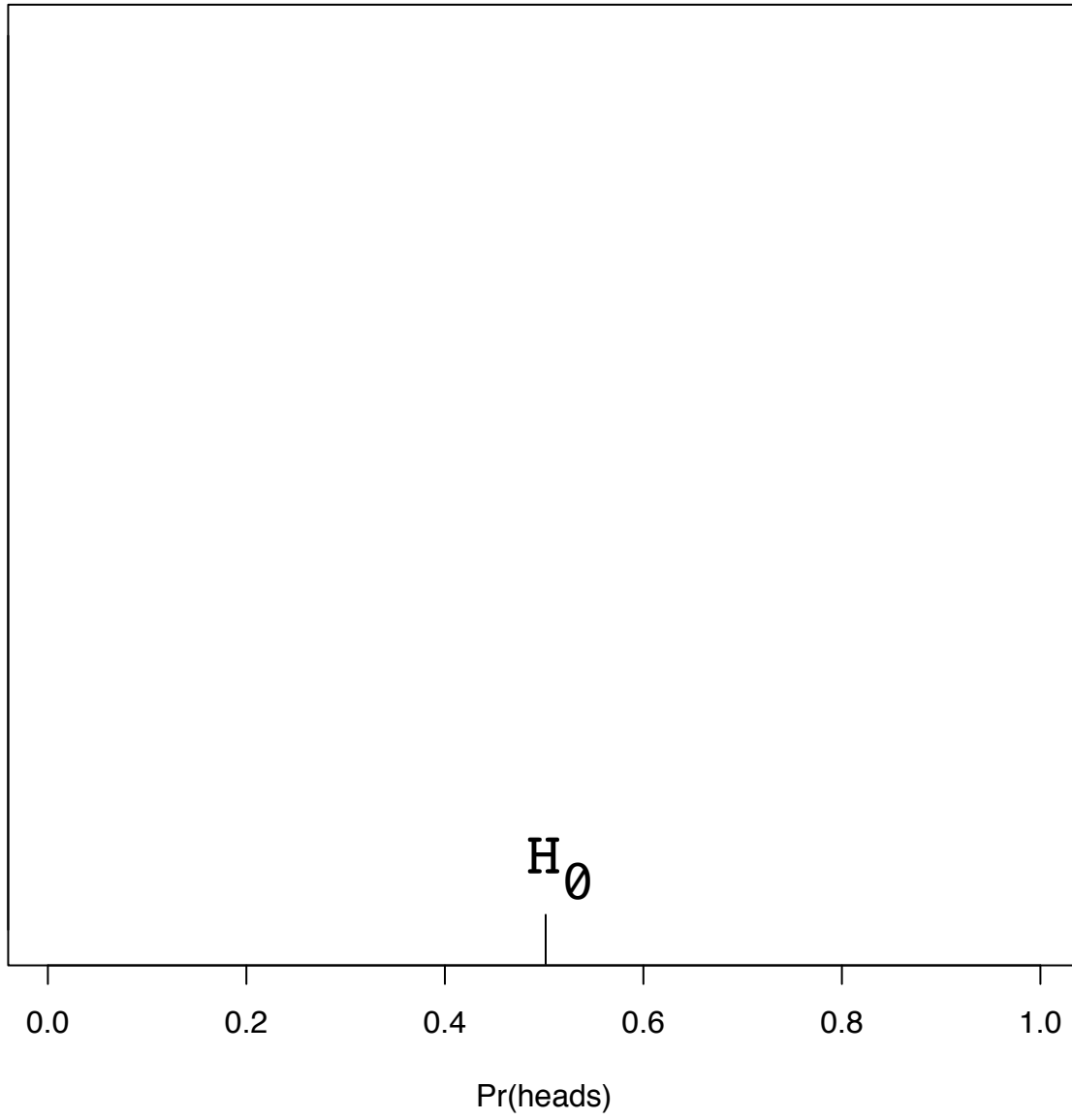
Frequentist hypothesis testing: coin flipping example

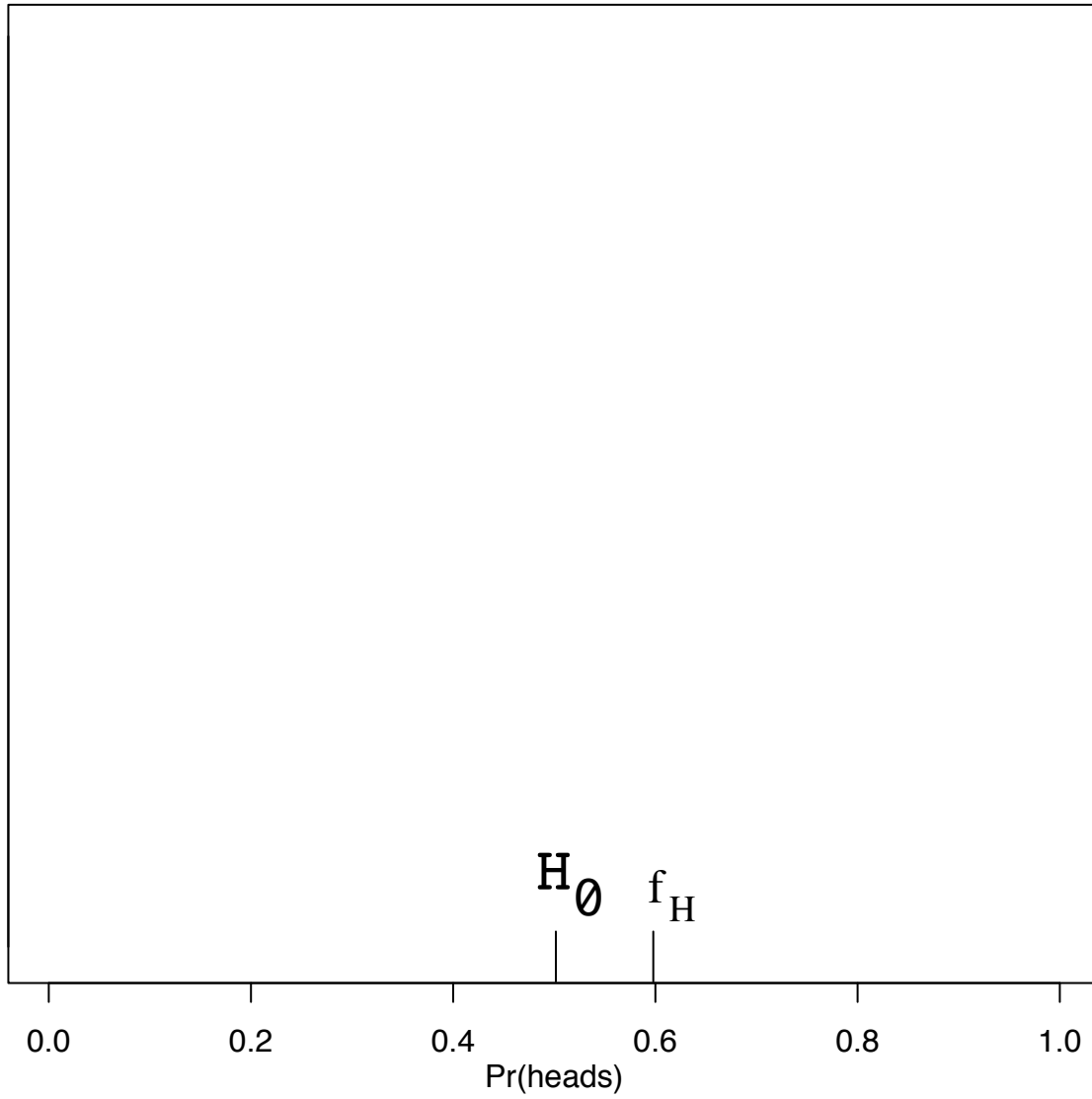
$N = 100$ and $h = 60$

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

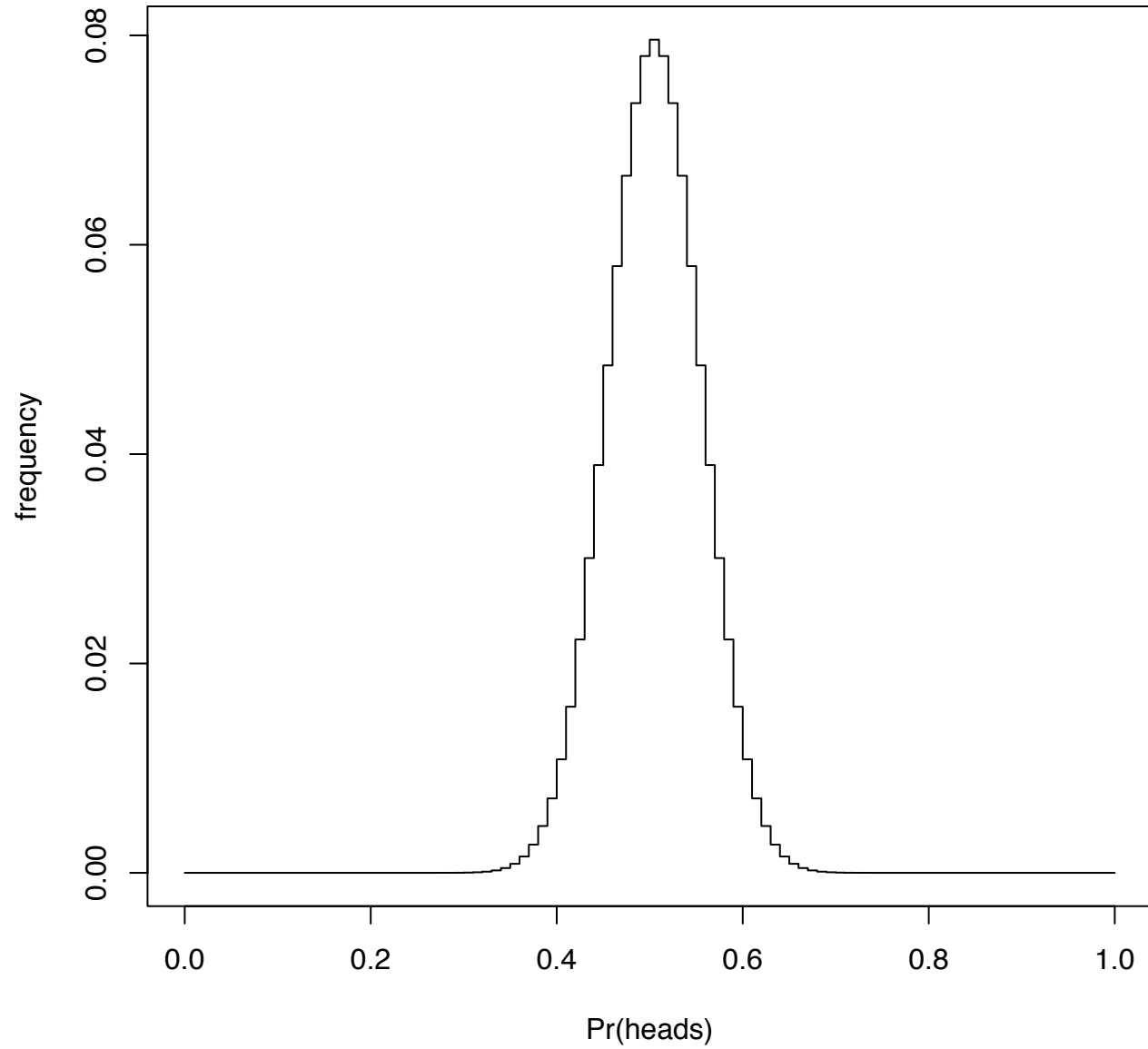
The “recipe” is:

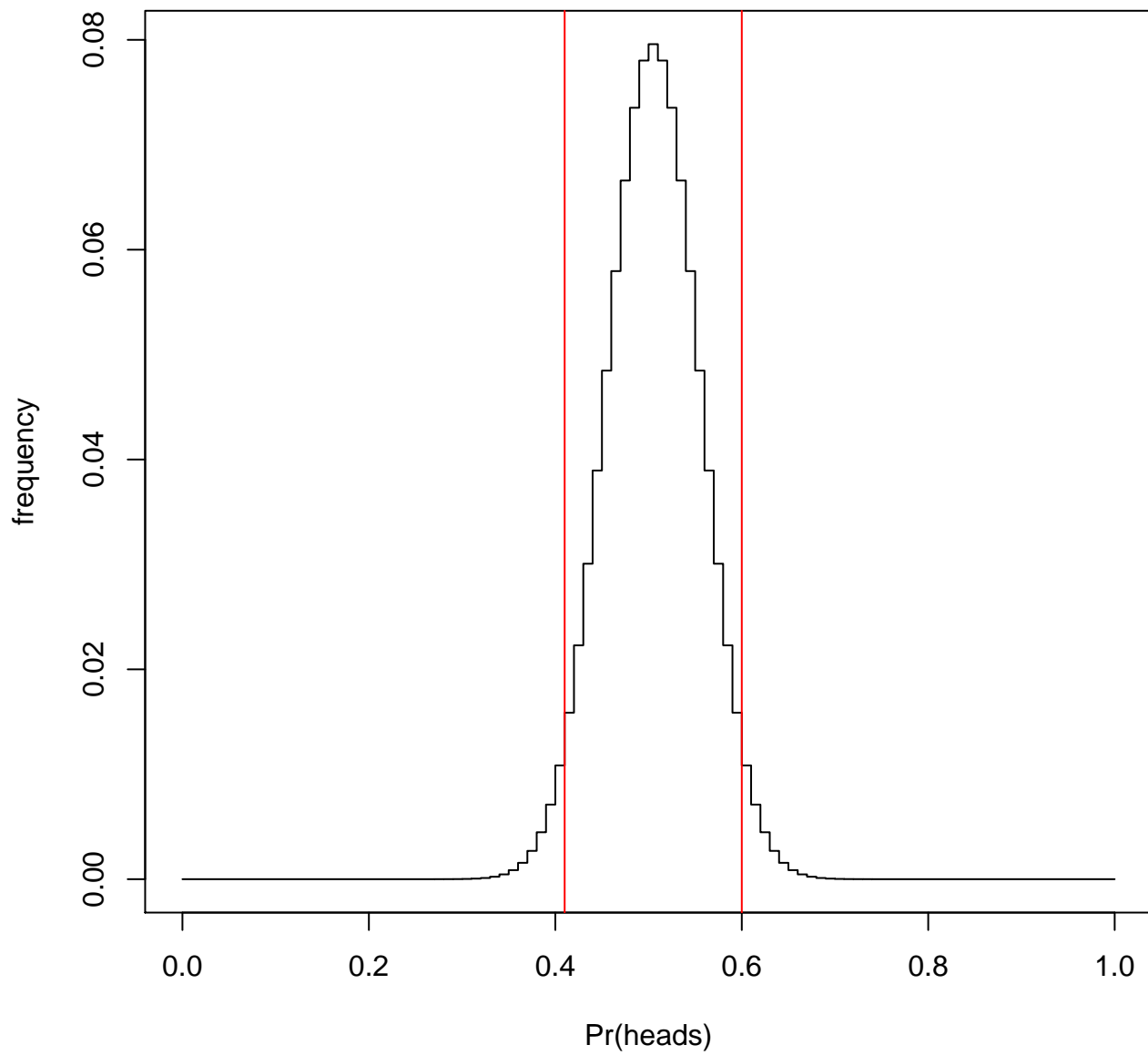
1. Formulate null (H_0) and alternative (H_A) hypotheses.
2. Choose an acceptable Type-I error rate (significance level)
3. Choose a test statistic: $f_H =$ fraction of heads in sample.
 $f_H = 0.6$
4. Characterize the null distribution of the test statistic
5. Calculate the P -value: The probability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

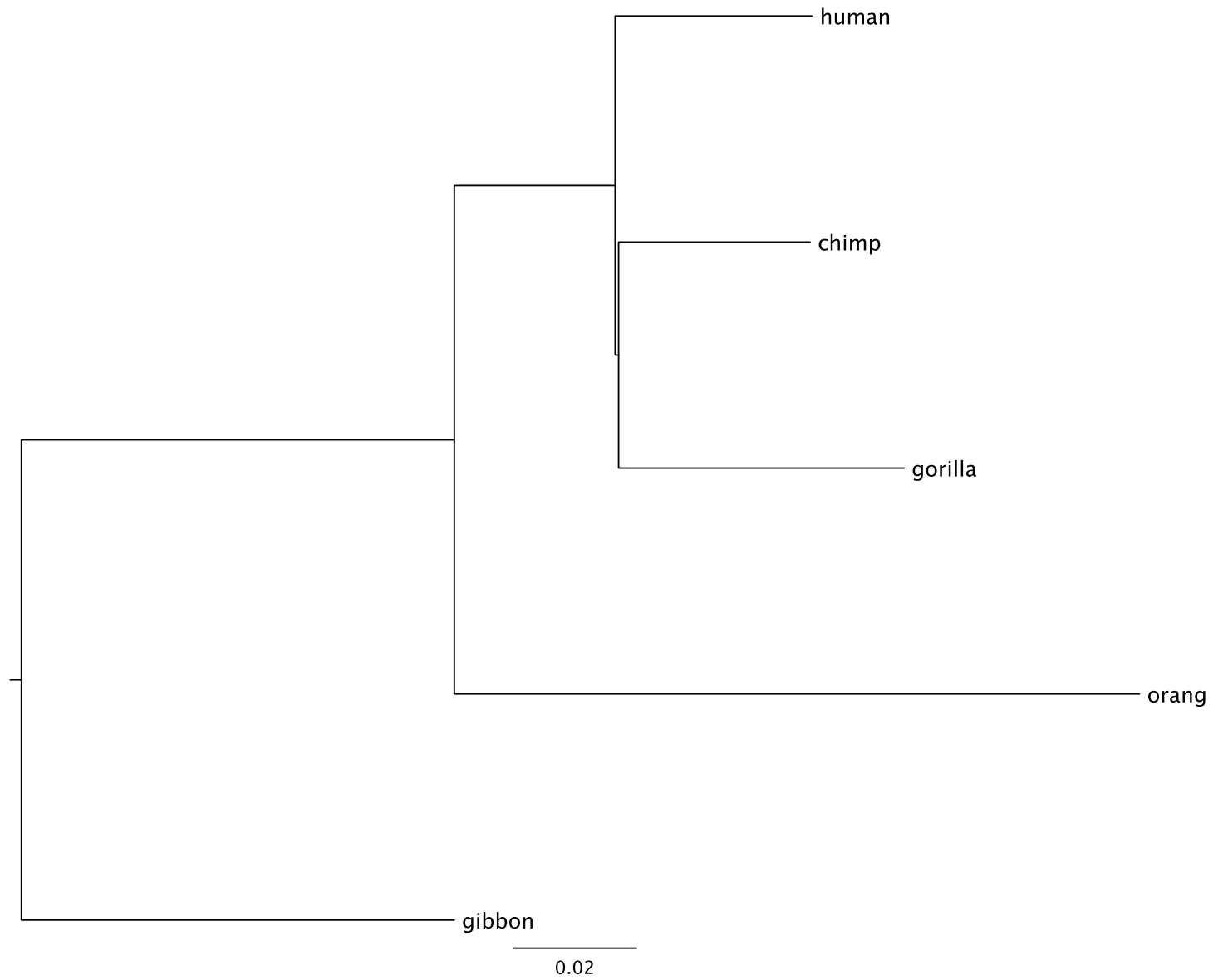
$$\delta(T_1, T_2 | X) = 2 [\ln L(T_1 | X) - \ln L(T_2 | X)]$$

Expectation under null:

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

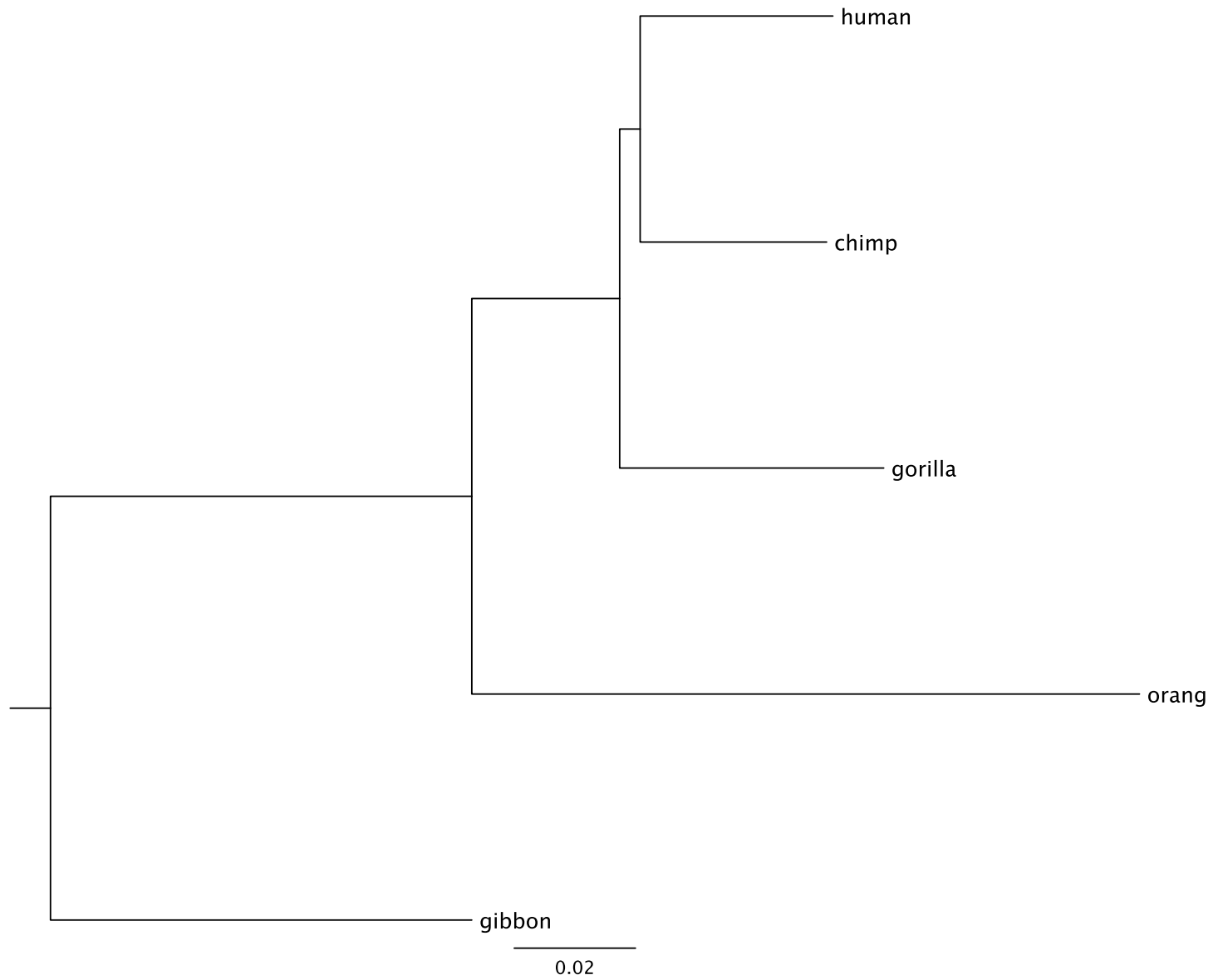
Using 3000 sites of mtDNA sequence for 5 primates

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

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

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

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

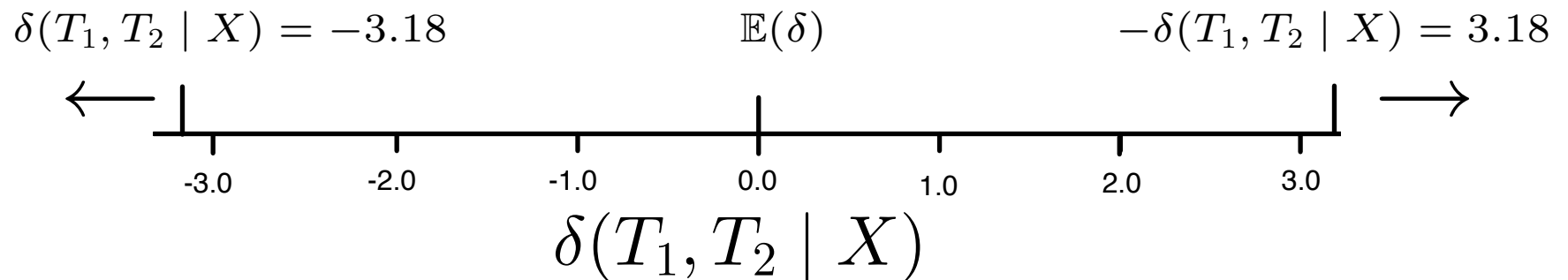
$$\mathbb{E}(\delta)$$



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

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

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



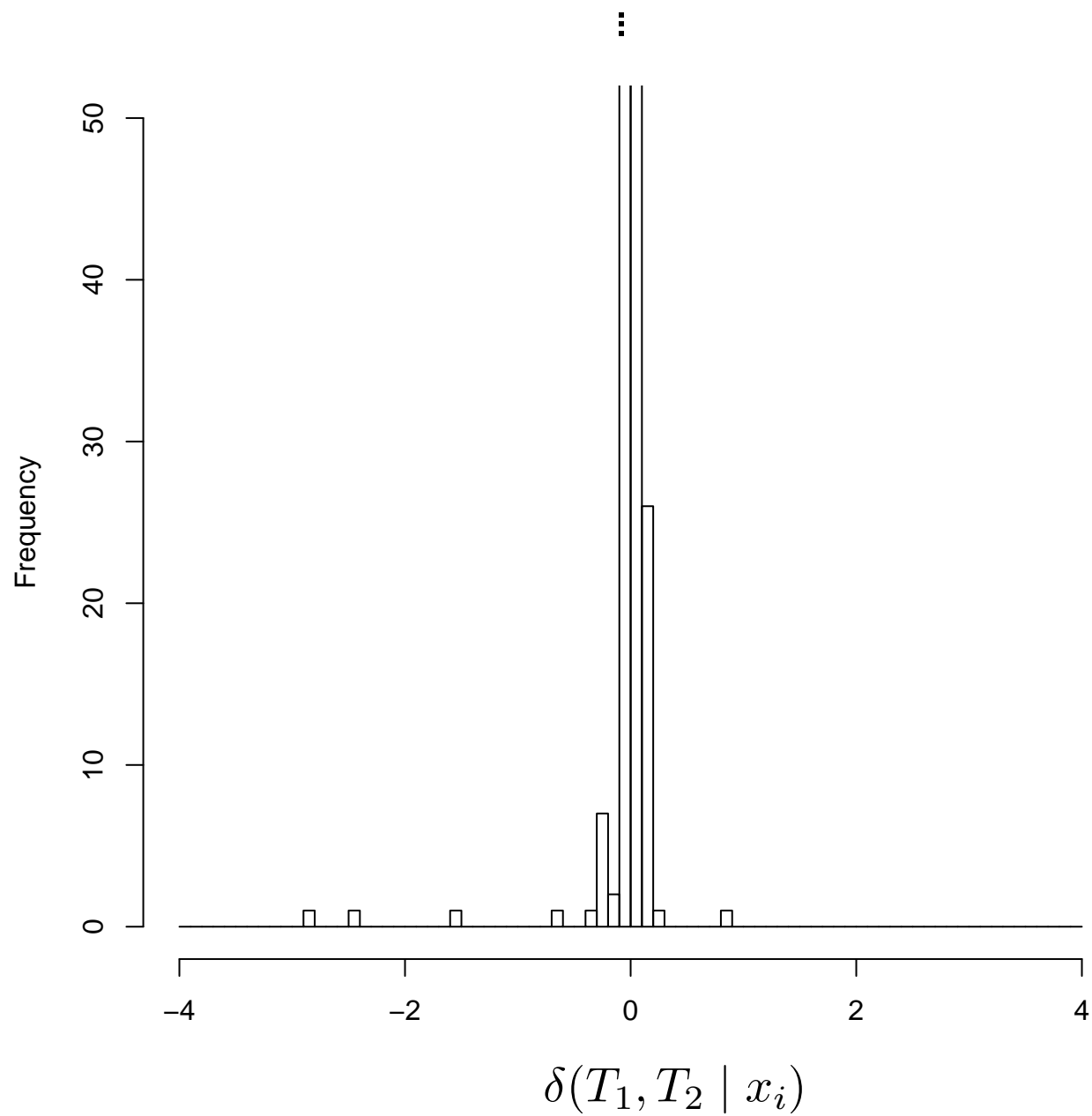
KH Test

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.

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

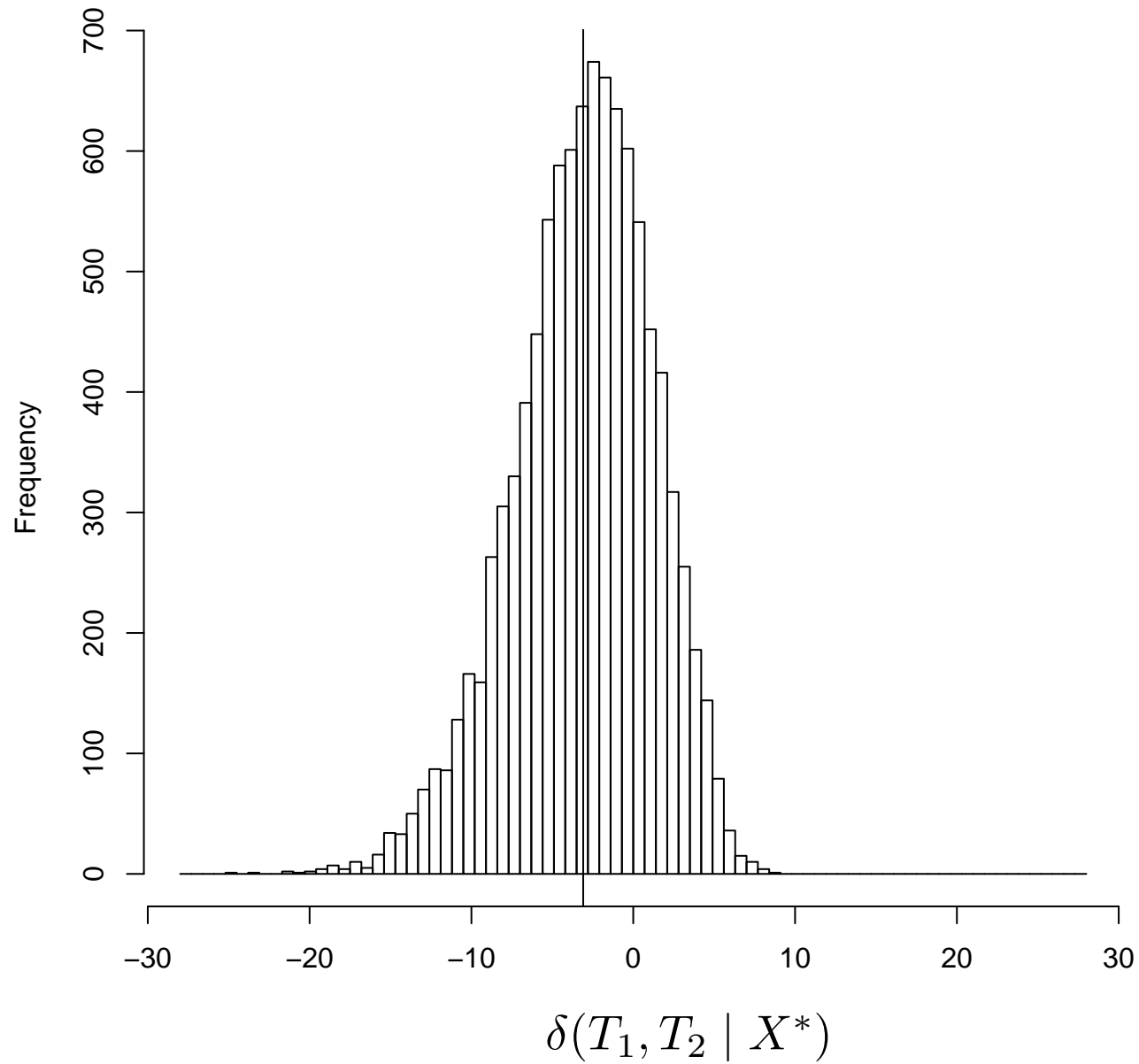


KH Test - the variance of $\delta(T_1, T_2 | X)$

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.

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.

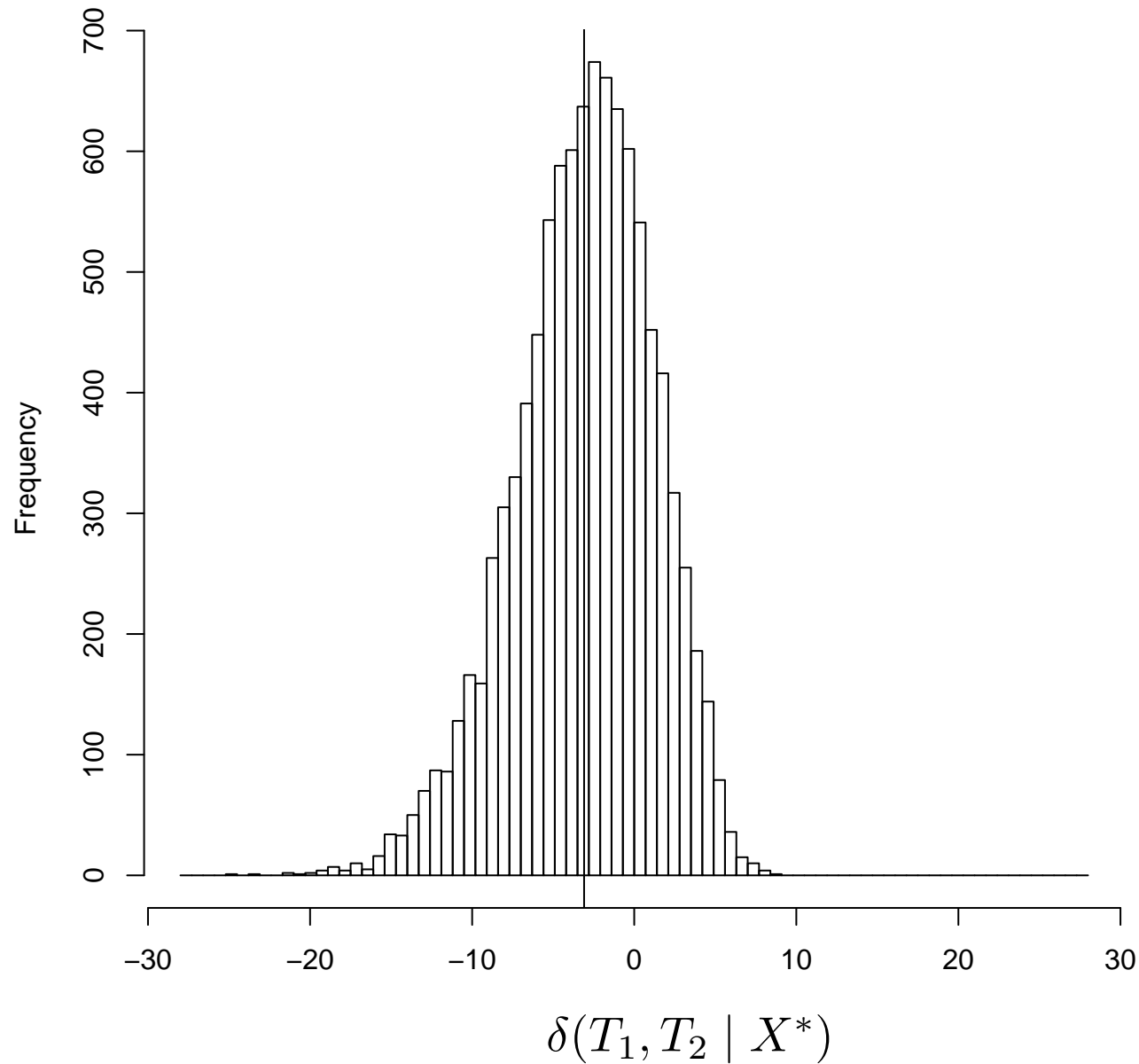
So, we can simply resample the site $\ln L$ values and sum them (rather than reoptimizing parameters).

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:

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

Bootstrapping gives us a reasonable guess of the variance under H_0

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

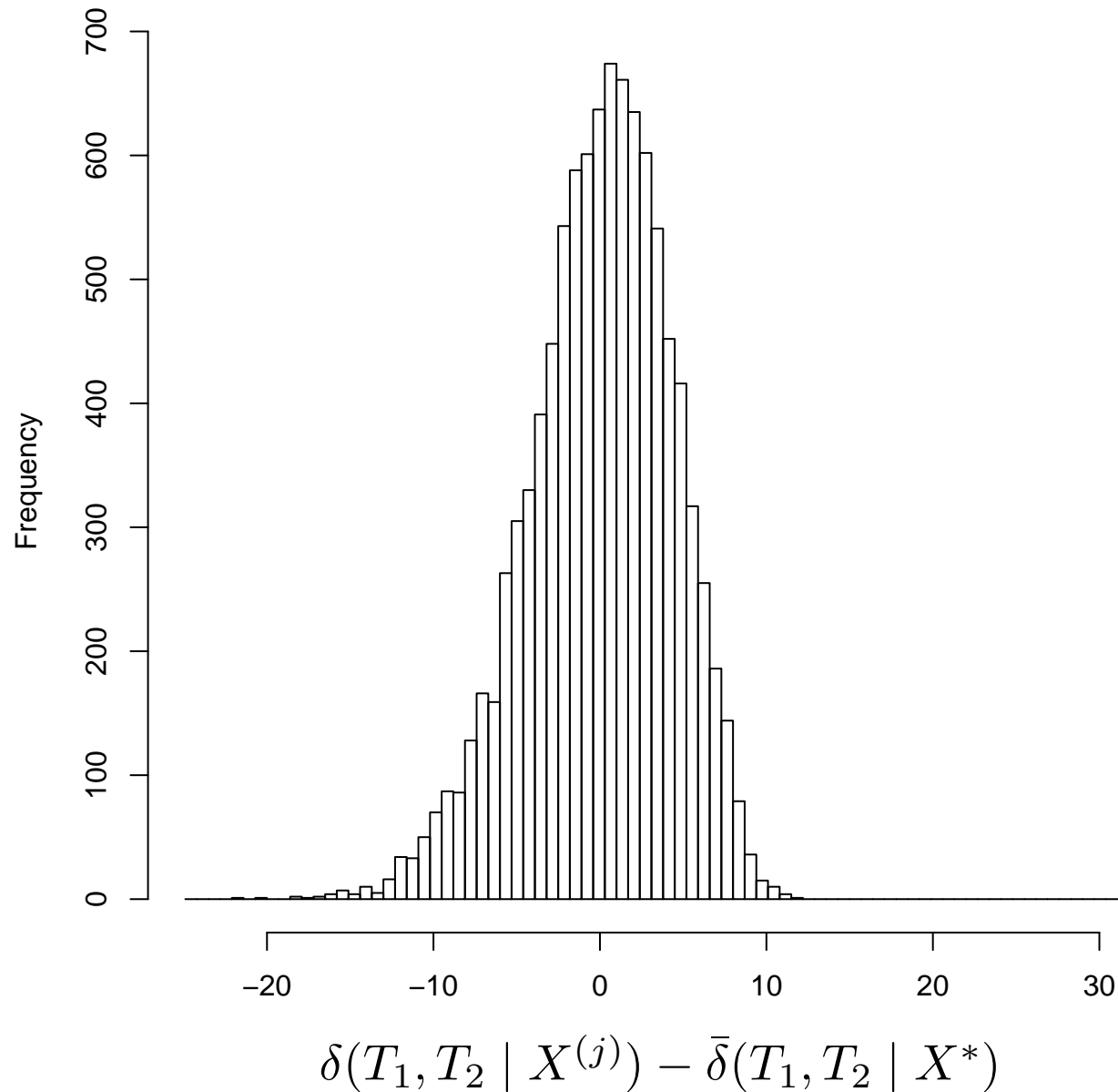
For each of the j bootstrap replicates, we treat

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

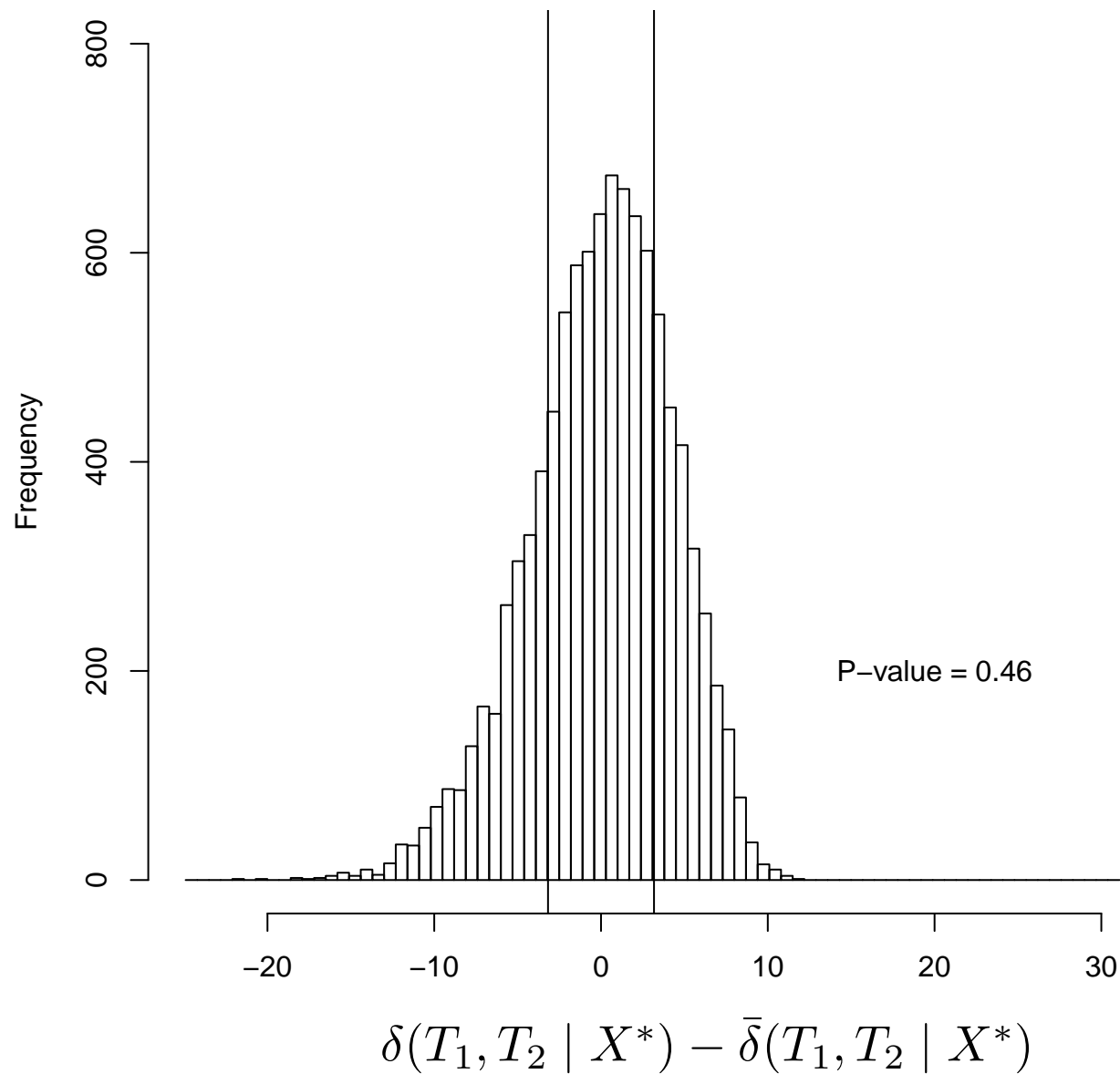
as draws from the null distribution.

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

for many (RELL) bootstrapped replicates of the data



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



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 , differs 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

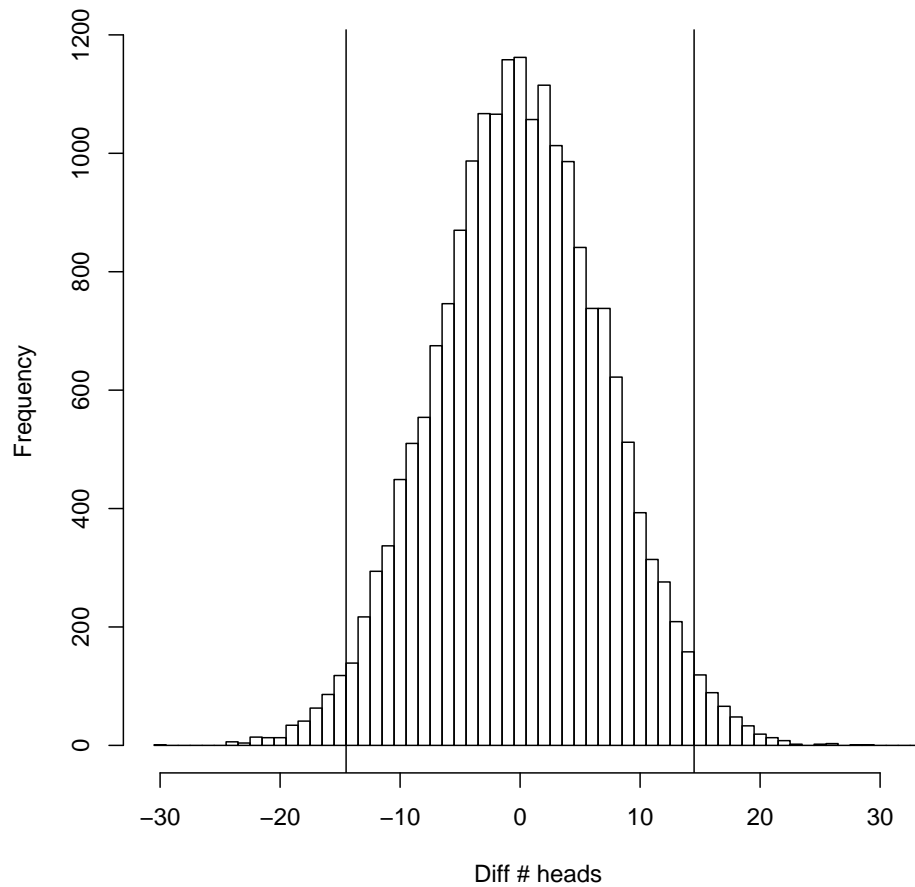
$$2 \left[\ln L(\hat{T}) - \ln L(T_1) \right] = 0$$

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.

$$h_1 - h_2$$

Null dist.: difference in # heads any two competitors

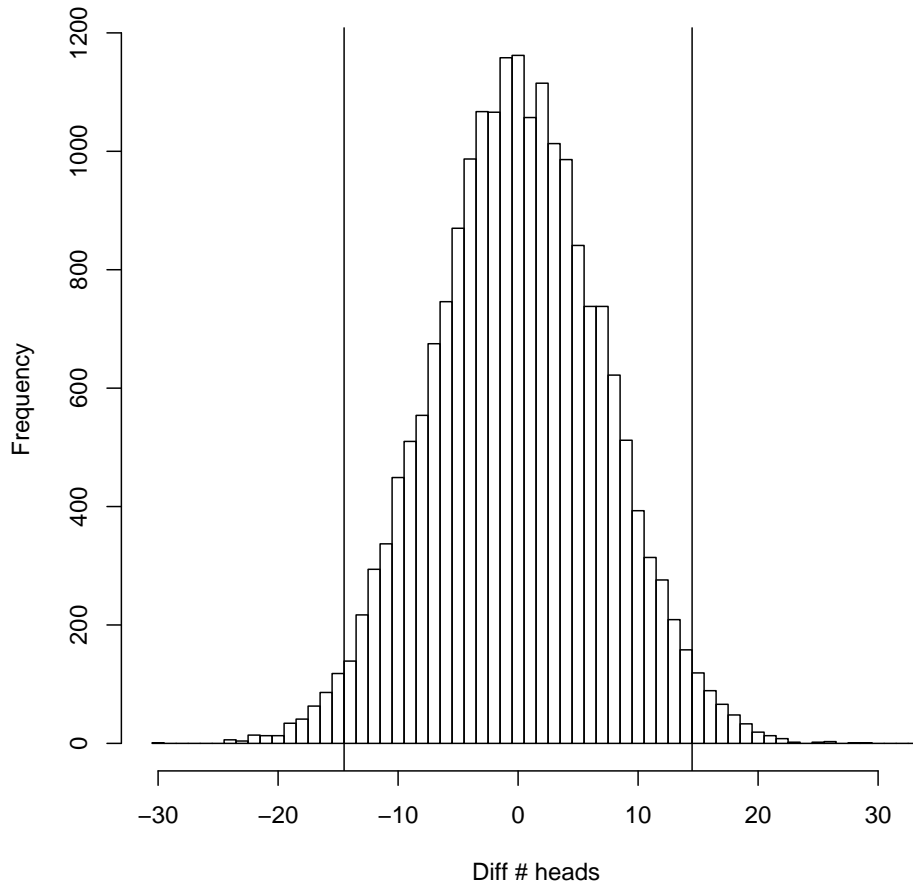


Experiment: 70 people each flip a fair coin 100 times and count # heads.

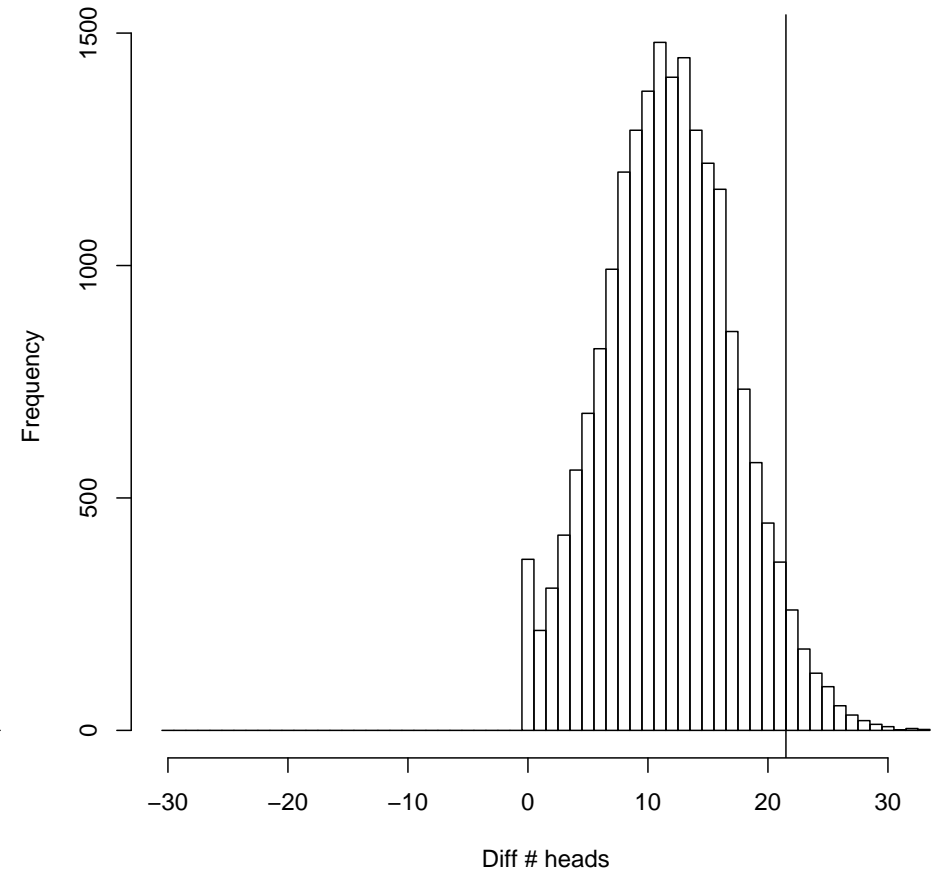
$$h_1 - h_2$$

$$\max(h) - h_1$$

Null dist.: difference in # heads any two competitors



Null dist.: difference highest – random competitor



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 trees¹

¹Because 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, $\ln \bar{L}(T_i | X^*)$, over all bootstrap replicates to center the bootstrapped collection of log-likelihoods:

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

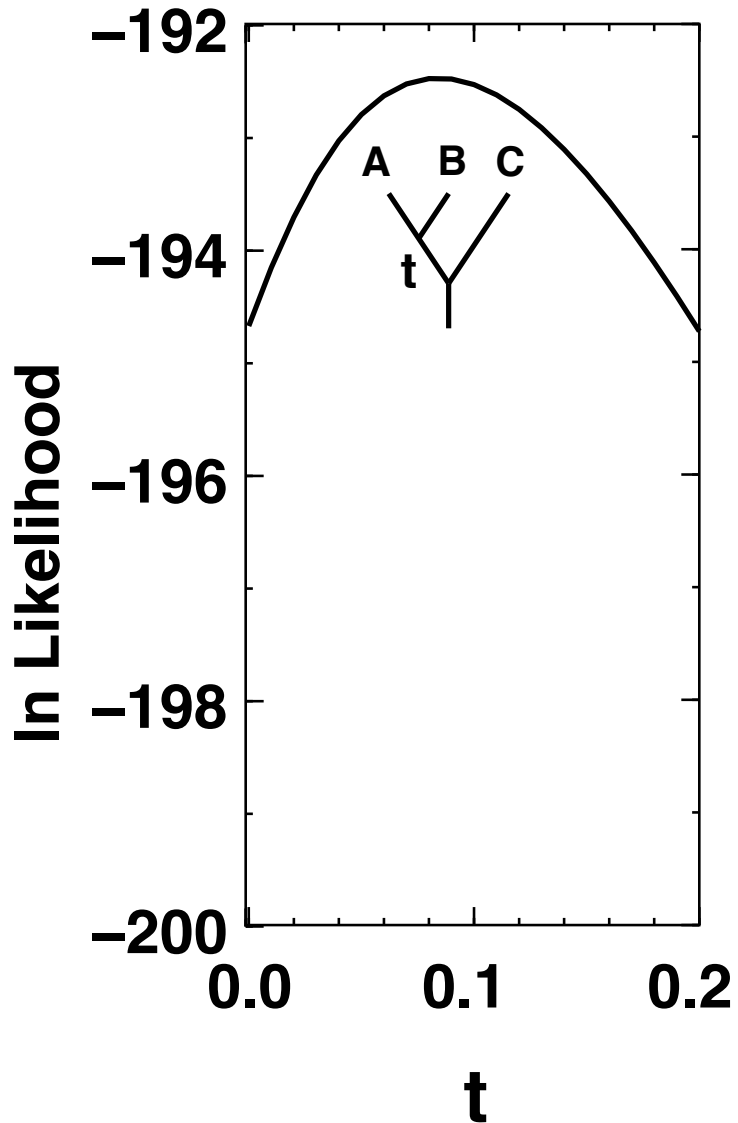
- For each bootstrap replicate, j , pick the highest value from the centered distributions (this mimics the selection bias):

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

- 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:

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

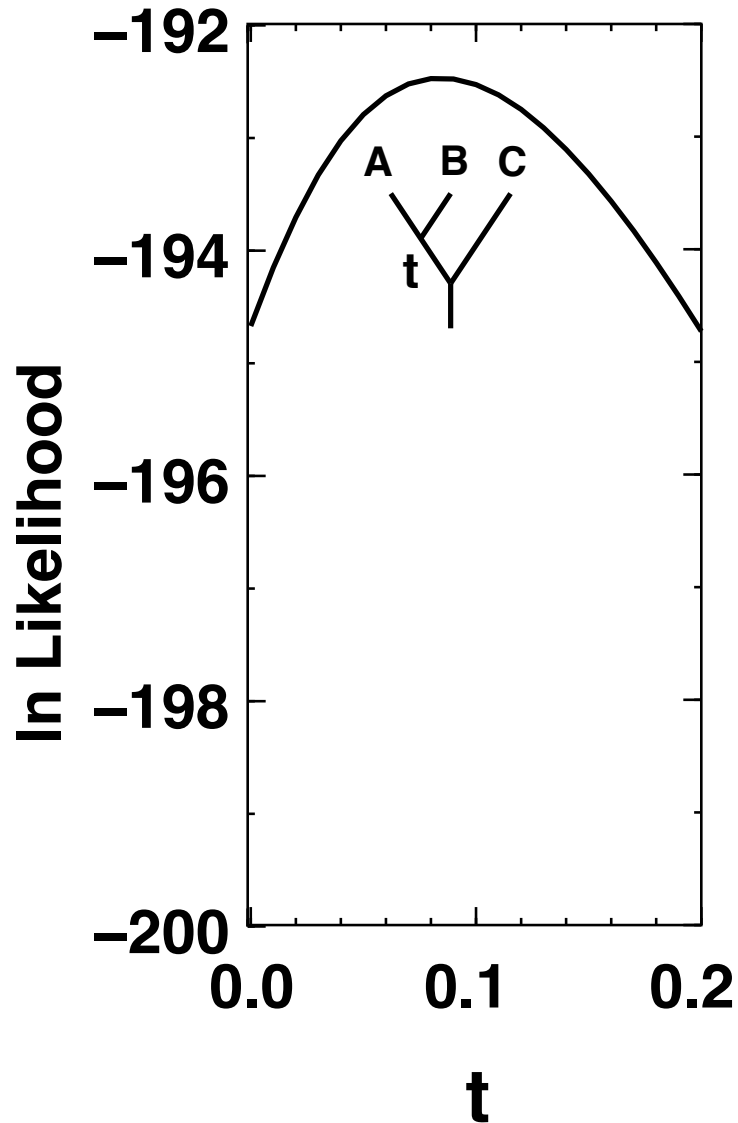
Can we test trees using the LRT?



1. Should we calculate the LRT as:
$$\delta_i = 2 [\ln L(t = \hat{t}, T_i | X) - \ln L(t = 0, T_i | X)]$$

2. And can we use the χ_1^2 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:
$$\delta_i = 2 [\ln L(t = \hat{t}, T_i | X) - \ln L(t = 0, T_i | X)]$$

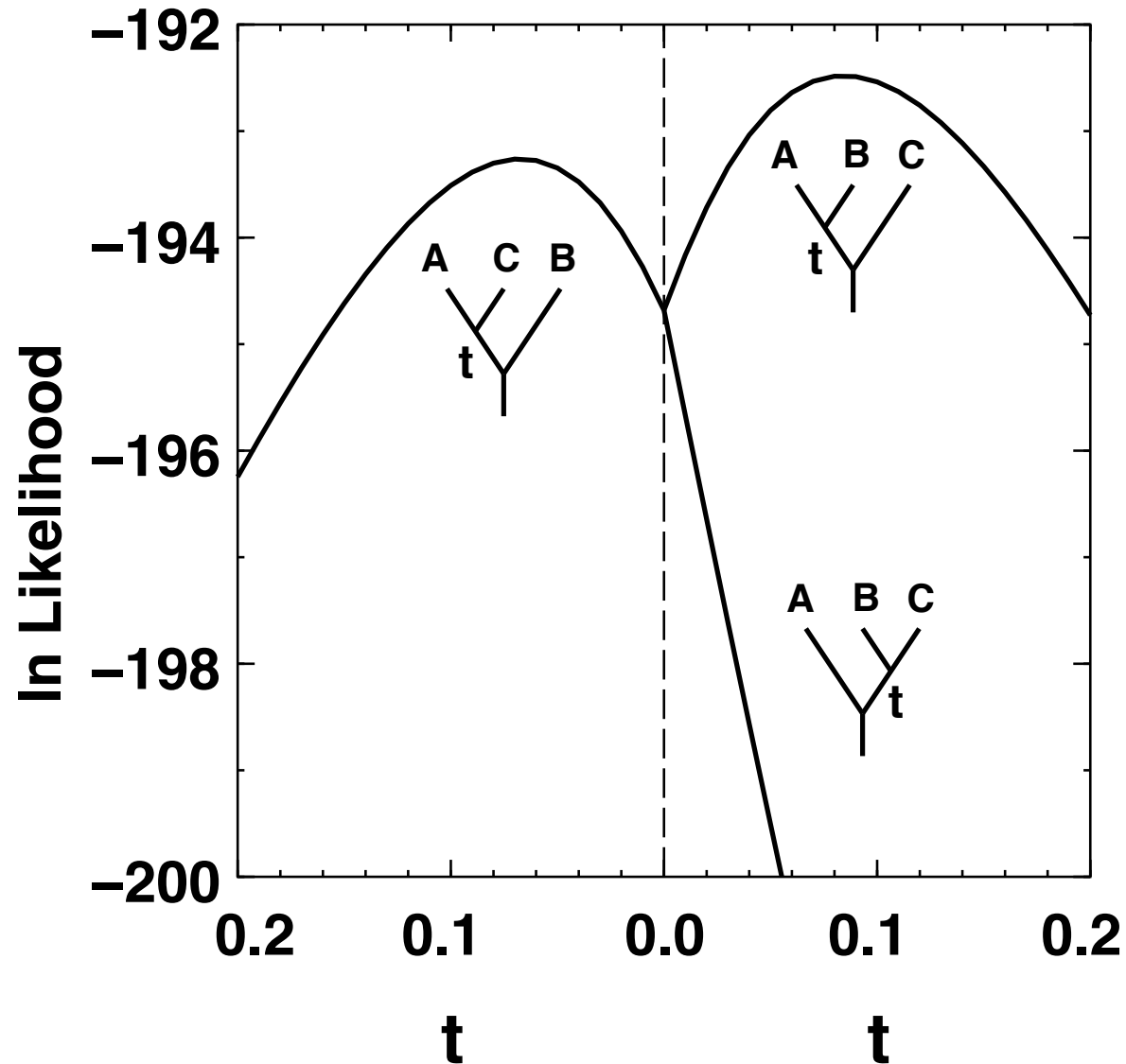
No. $t = 0$ might not yield the best alternative $\ln L$

2. And can we use the χ_1^2 distribution to get the critical value for δ ?

No. Constraining parameters at boundaries leads to a mixture such as: $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$

See Ota et al. (2000).

Can we test trees using the LRT?



No, tree hypotheses are not nested!

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

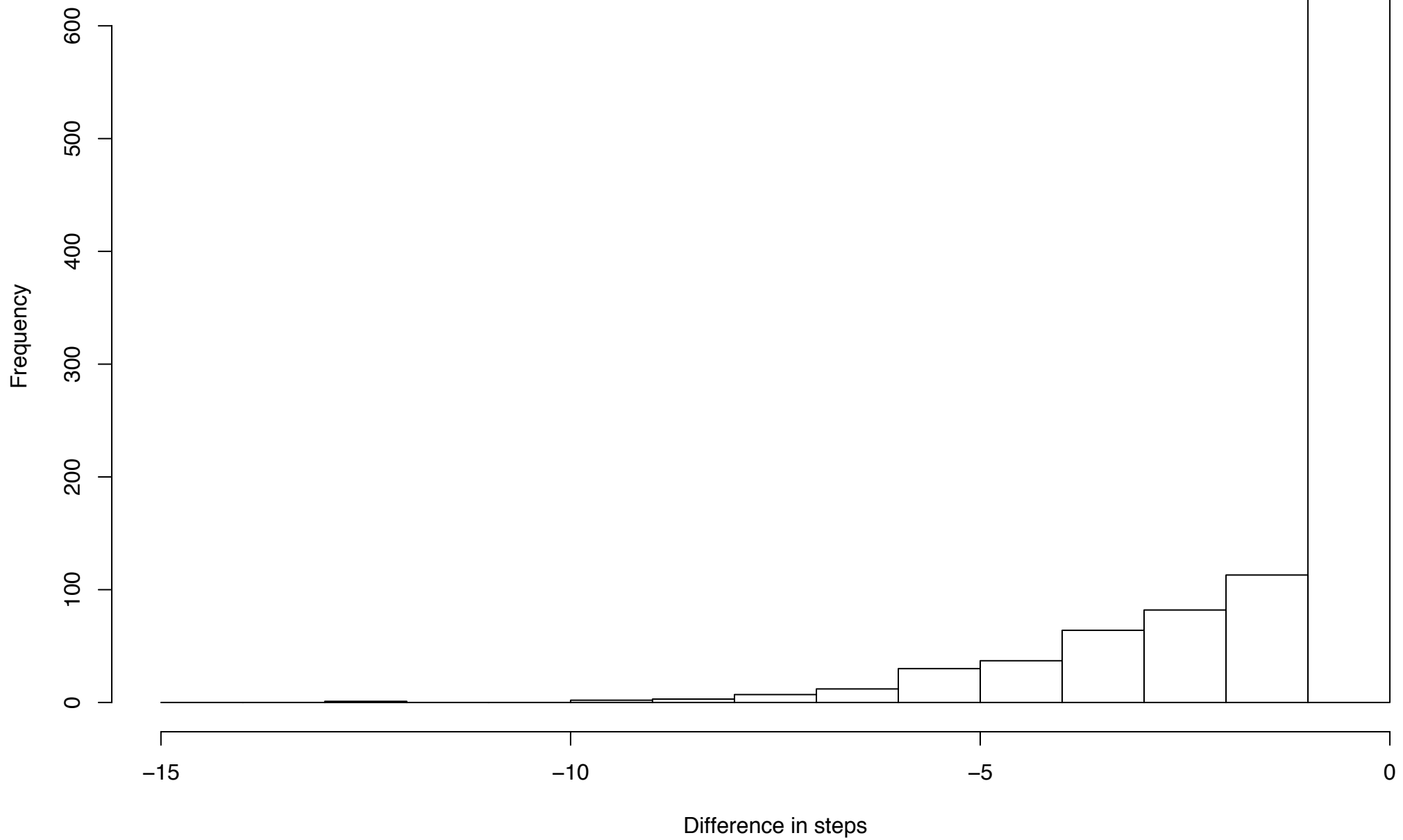
Parametric bootstrapping

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

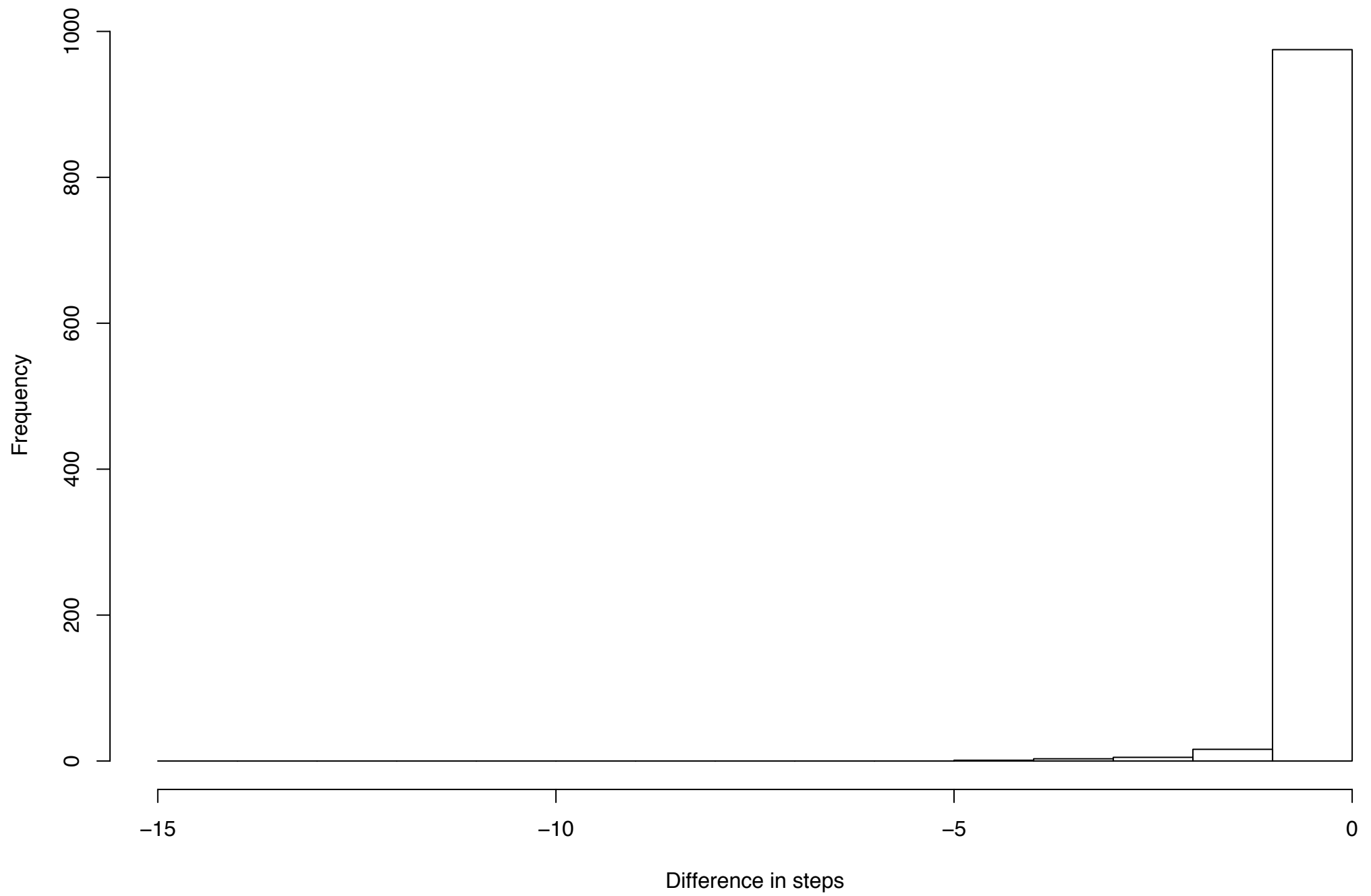
Huelsenbeck et al. (1996) describes how to use the approach as a test for monophyly.

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

Null distribution of the difference in number of steps under GTR+I+G

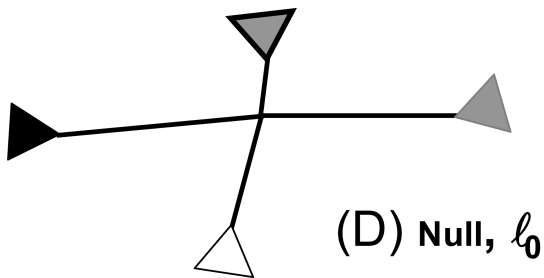
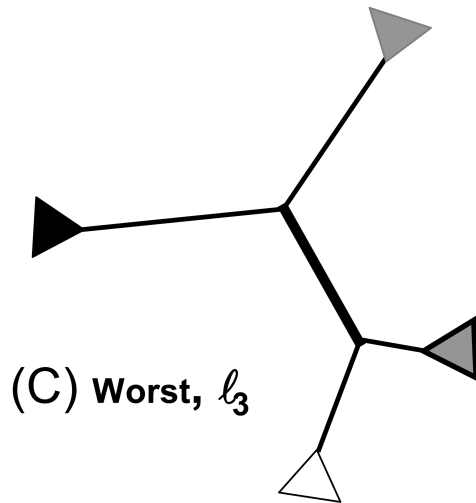
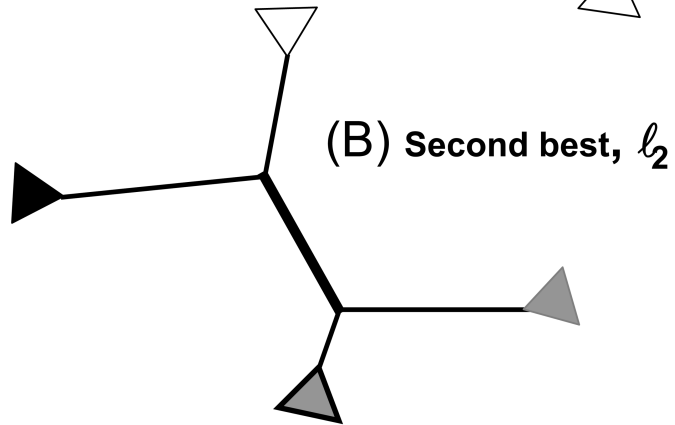
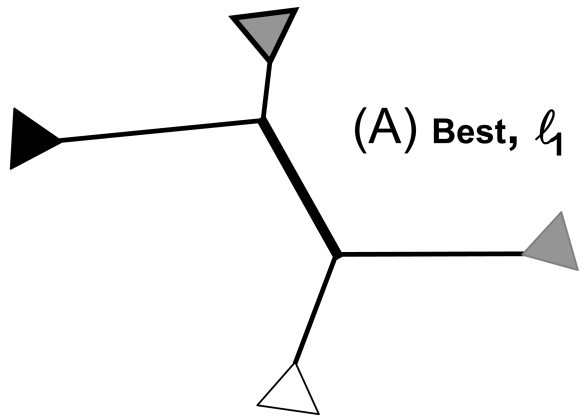


Null distribution of the difference in number of steps under JC



aLRT of Anisimova and Gascuel (2006)

- For a **branch** j , calculate δ_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
- They 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.



(A) Best, l_1

(B) Second best, l_2

(C) Worst, l_3

(D) Null, l_0

$$aLRT = 2 [\ln l_1 - \ln L(T_2 | X)]$$

$$l_1 = L(T_1 | X)$$

Image from Anisimova and Gascuel (2006)

aBayes Anisimova et al. (2011)

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

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)

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

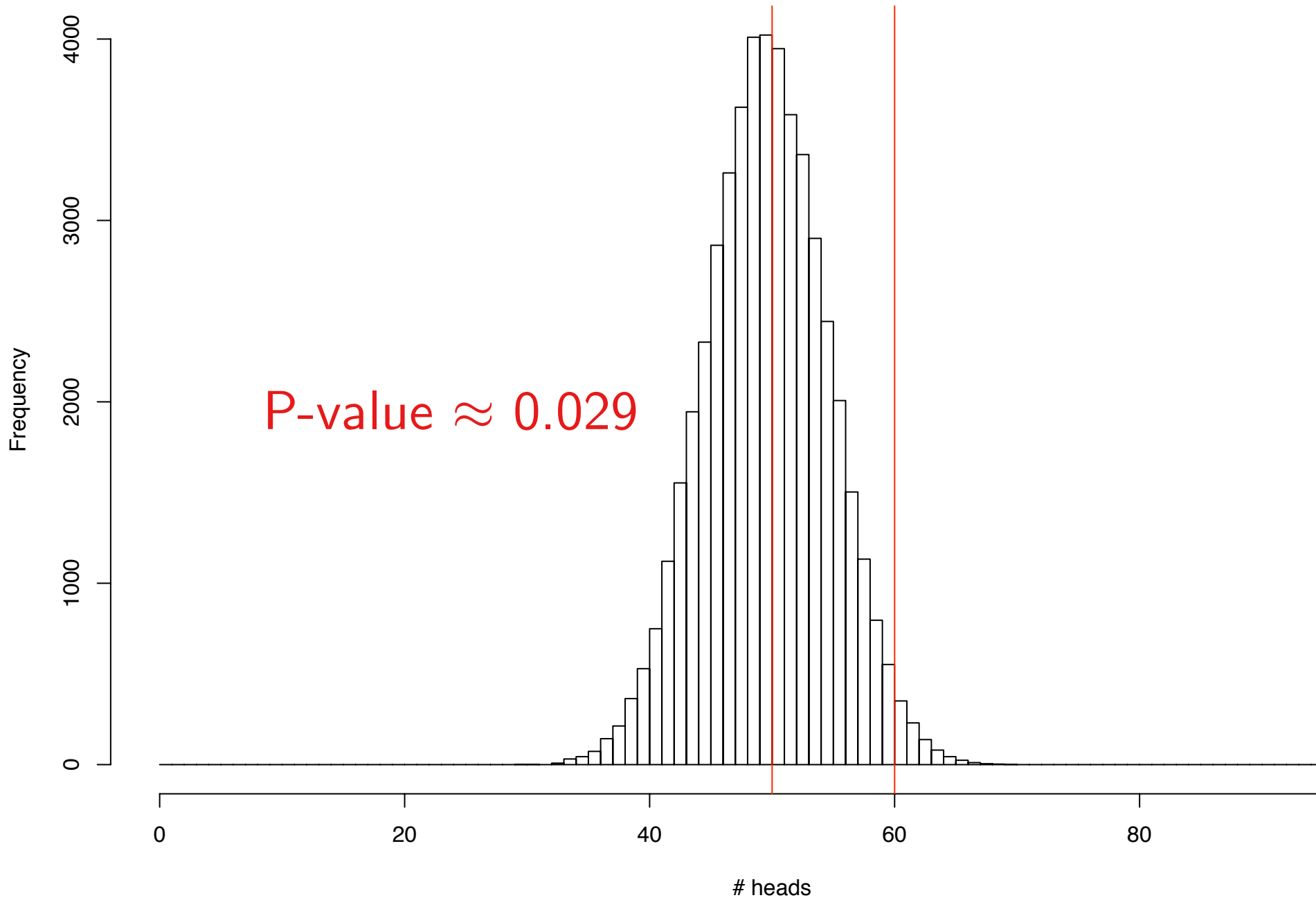
coin flipping (yet again)

$N = 100$ 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



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

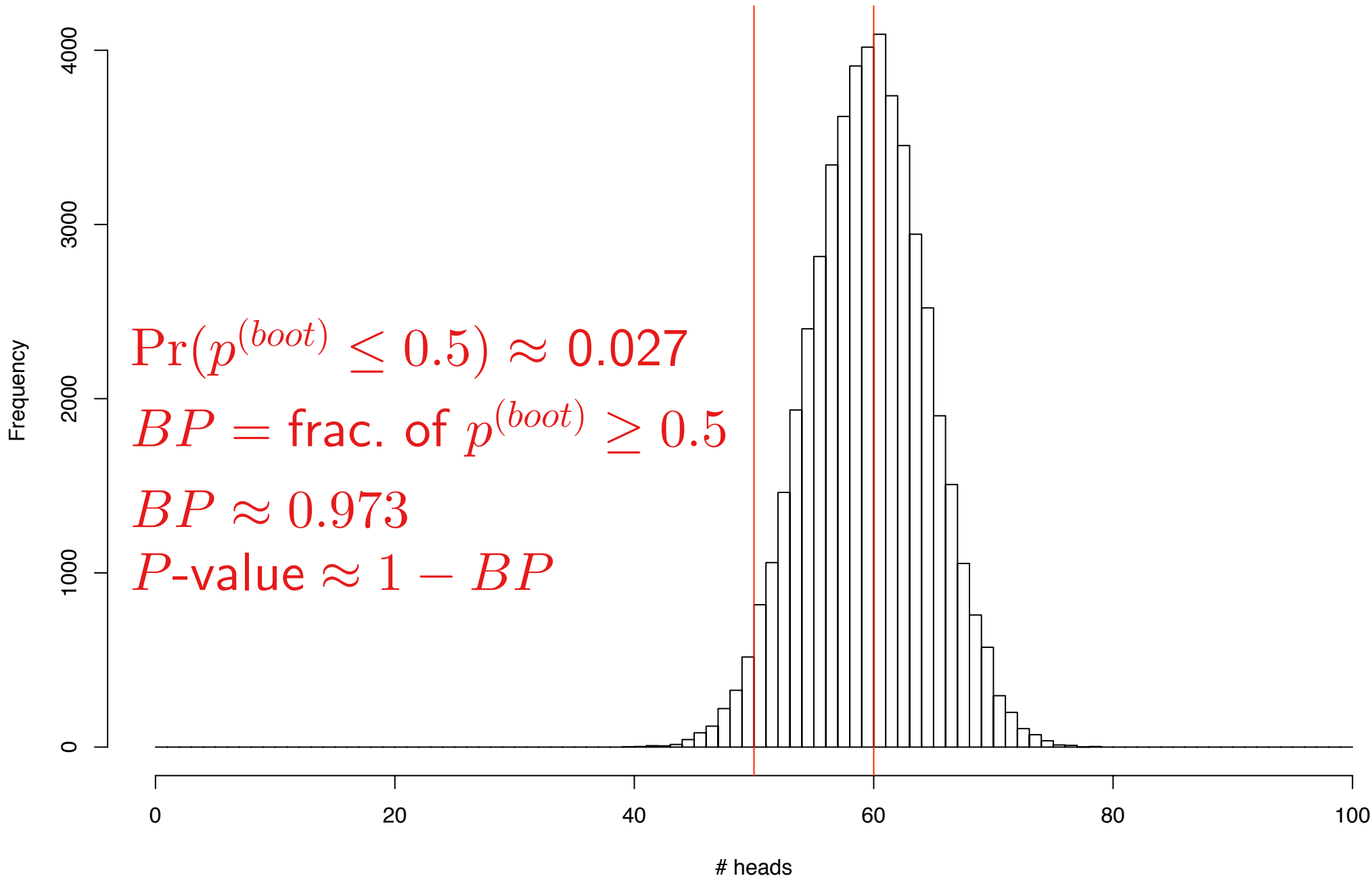
It can also give a tail probability for $\Pr(f_H^{(boot)} \leq 0.5)$

Amazingly (for many applications):

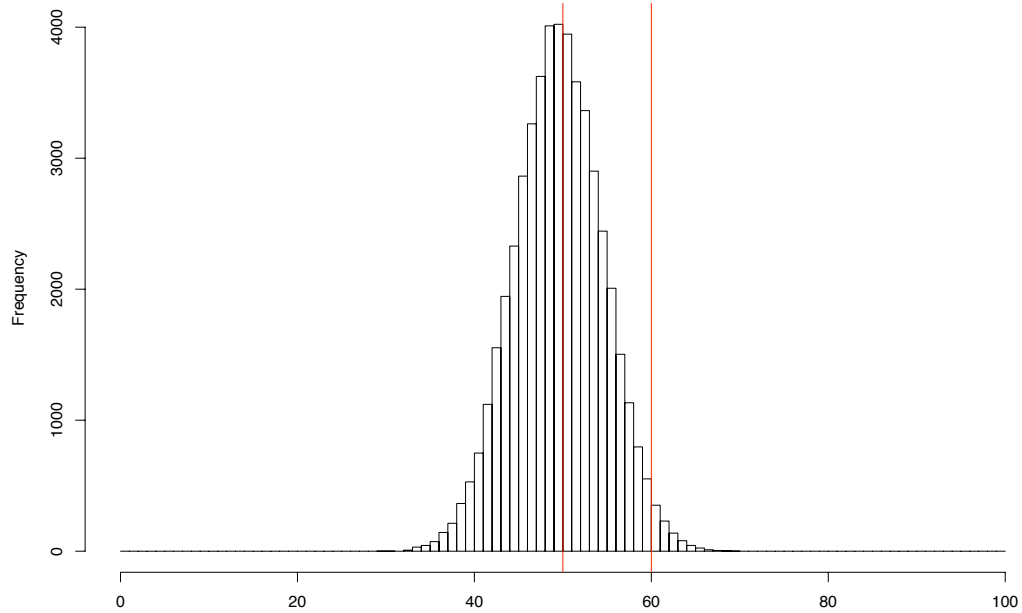
$$\Pr(\hat{f}_H \geq 0.6 \mid \text{null is true}) \approx \Pr(f_H^{(boot)} \leq 0.5)$$

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

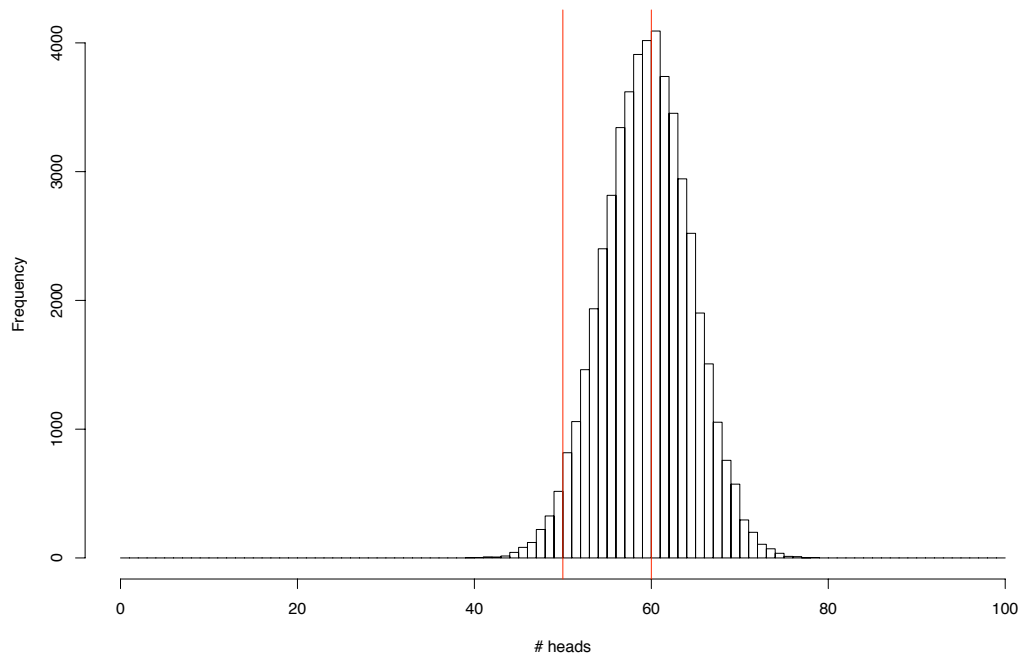
Distribution of the # heads in bootstrap resampled datasets



A simulation of the null distribution of the # heads

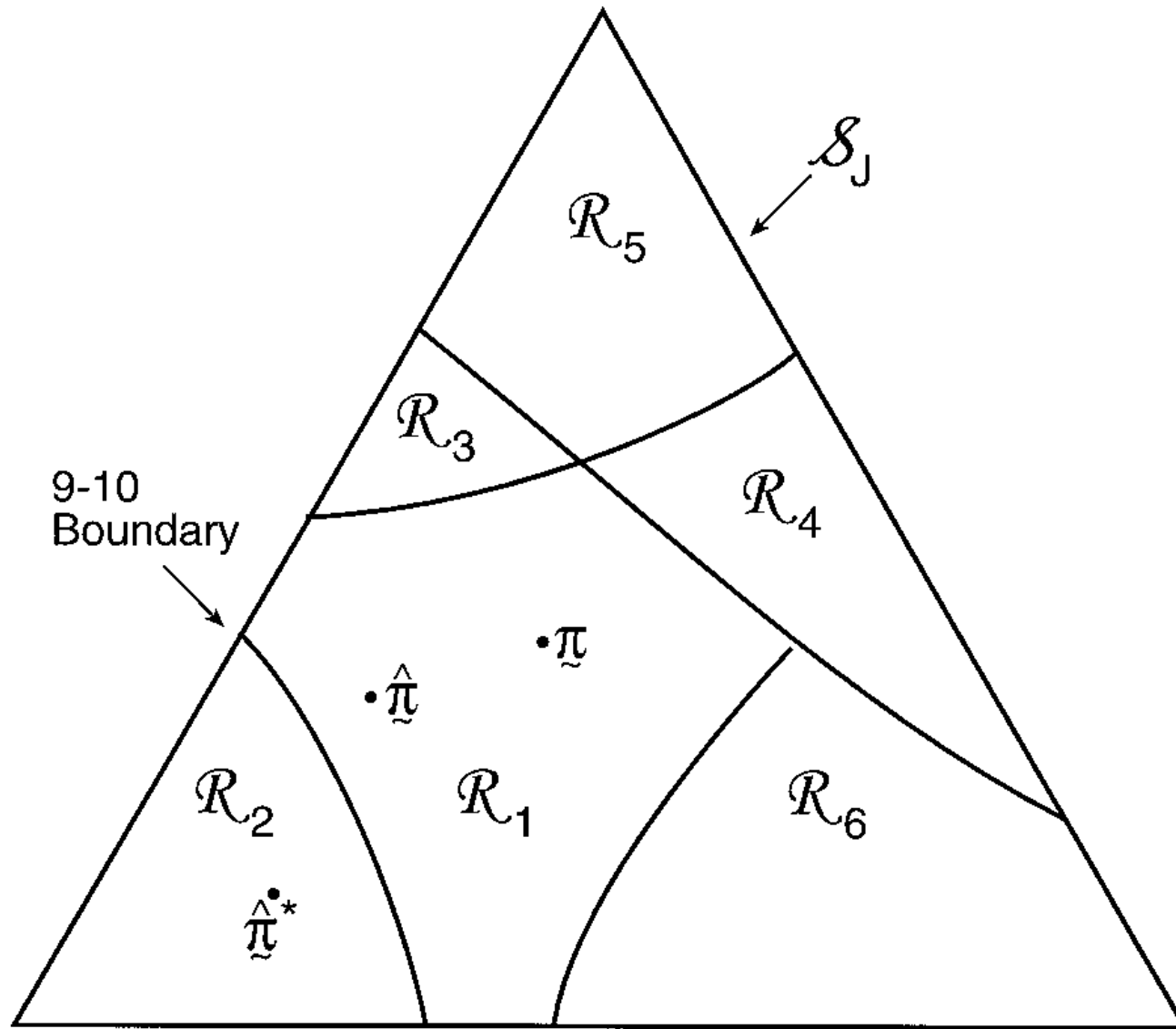


Distribution of the # heads in bootstrap resampled datasets

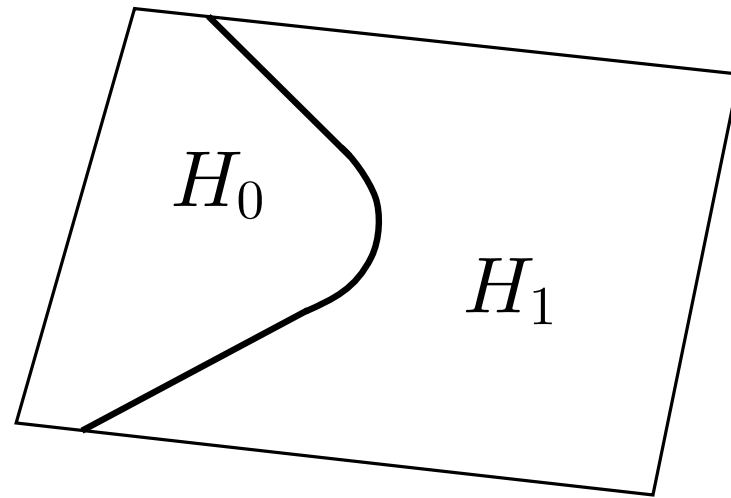
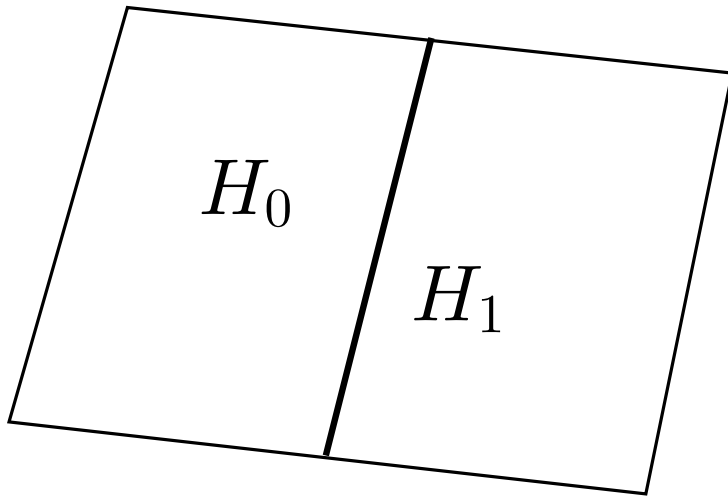


- When you decide between trees, the boundaries between tree hypotheses can be curved
 - When the boundary of the hypothesis space is curved, 1 - BP can be a poor approximation of the P -value.
- Efron et al. (1996)

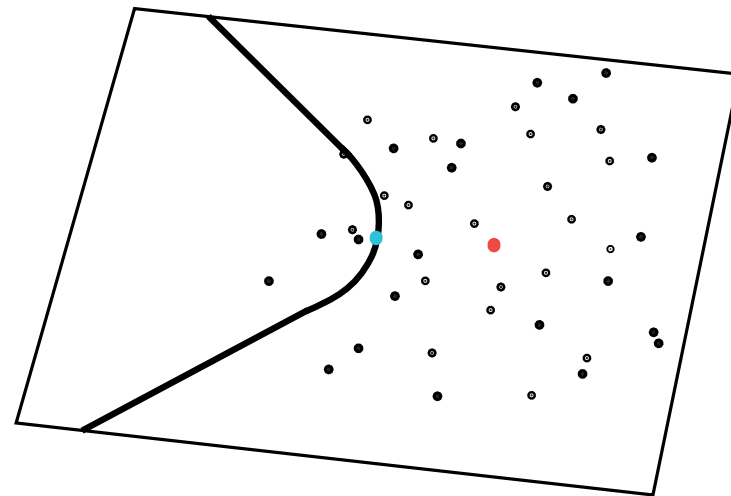
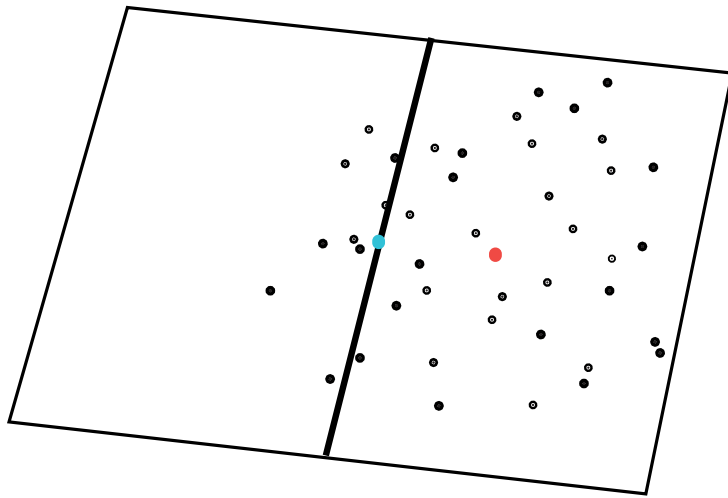
Efron et al. (1996) view of tree space



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.

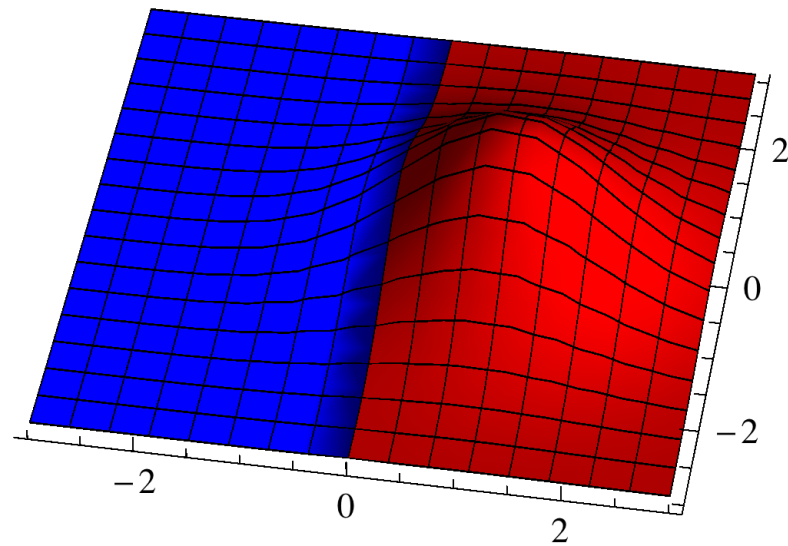
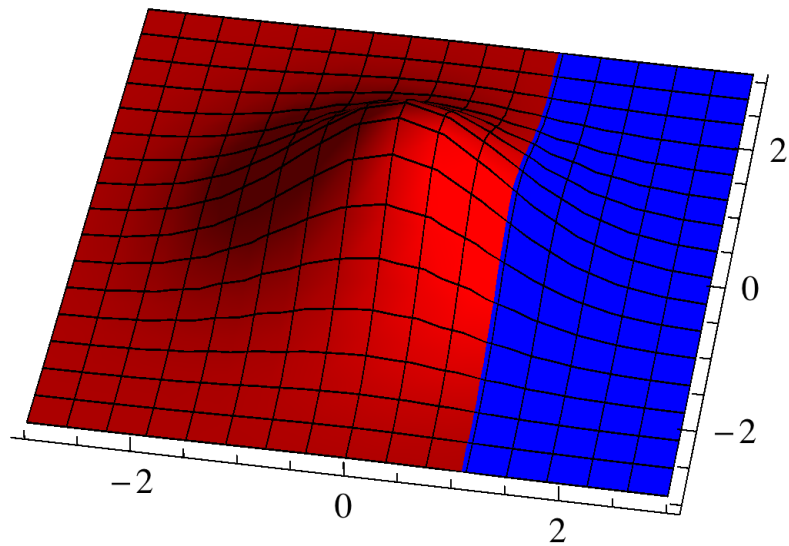


In the straight border case, symmetry implies that:

The actual P -value (blue region)

$$\approx 1 - BP$$

($1 - BP$ is the blue below)

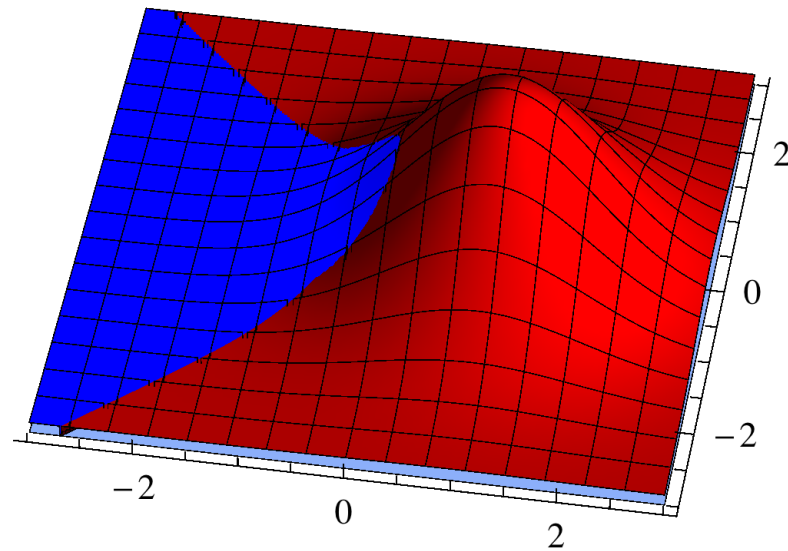
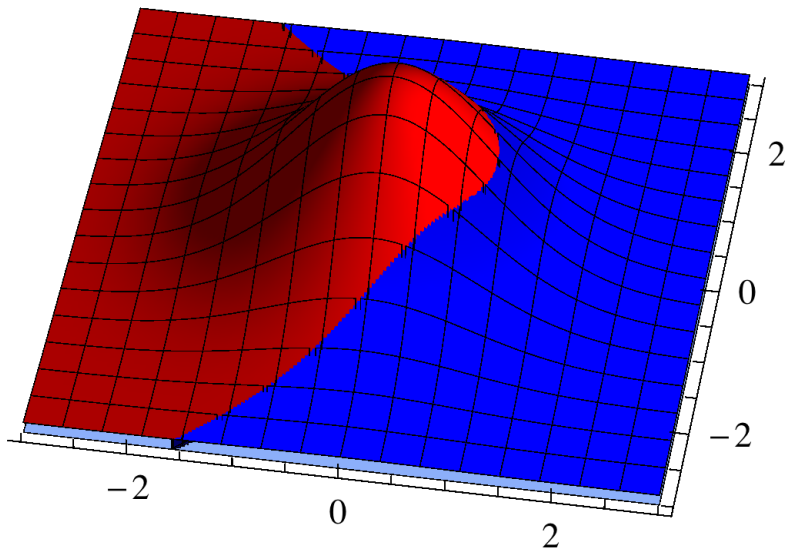


In the curved border case, the symmetry breaks down:

The actual P -value (blue region)

$$\neq 1 - BP$$

($1 - BP$ is the blue below)

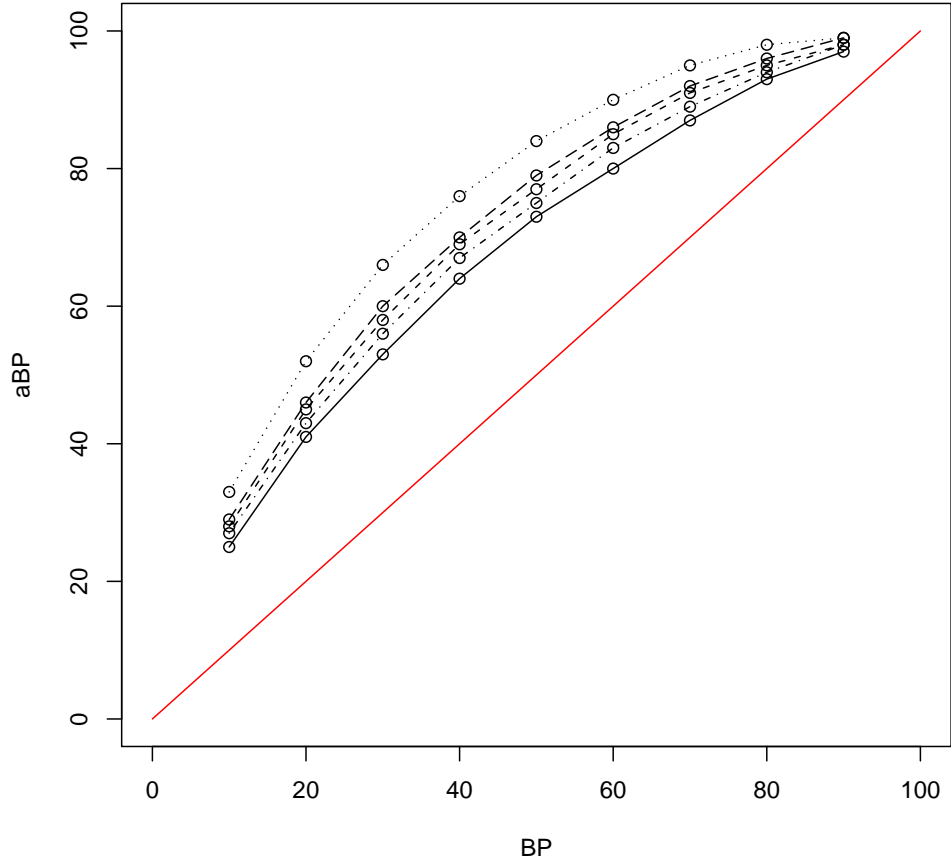


- Efron et al. (1996) proposed a computationally expensive multi-level bootstrap (which has not been widely used).
- Shimodaira (2002) used the same theoretical framework to devise a (more feasible) Approximately Unbiased (AU) test of topologies.
 - Multiple scales of bootstrap resampling (80% of characters, 90%, 100%, 110%...) are used to detect and correct for curvature of the boundary.
 - Implemented in the new versions of PAUP*

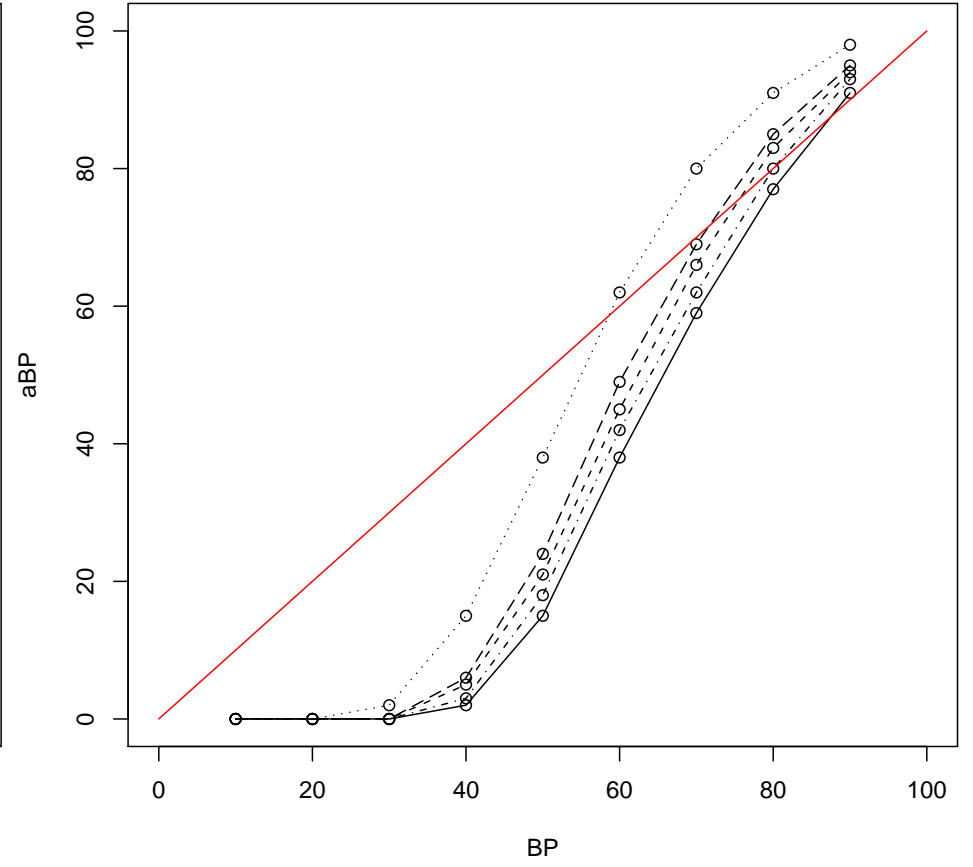
Susko (2010) adjusted BP – aBP

- 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 with selection bias correction for each BP (5 model conditions)



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.

Summary - Part 2

A (very) wide variety of tests differ by:

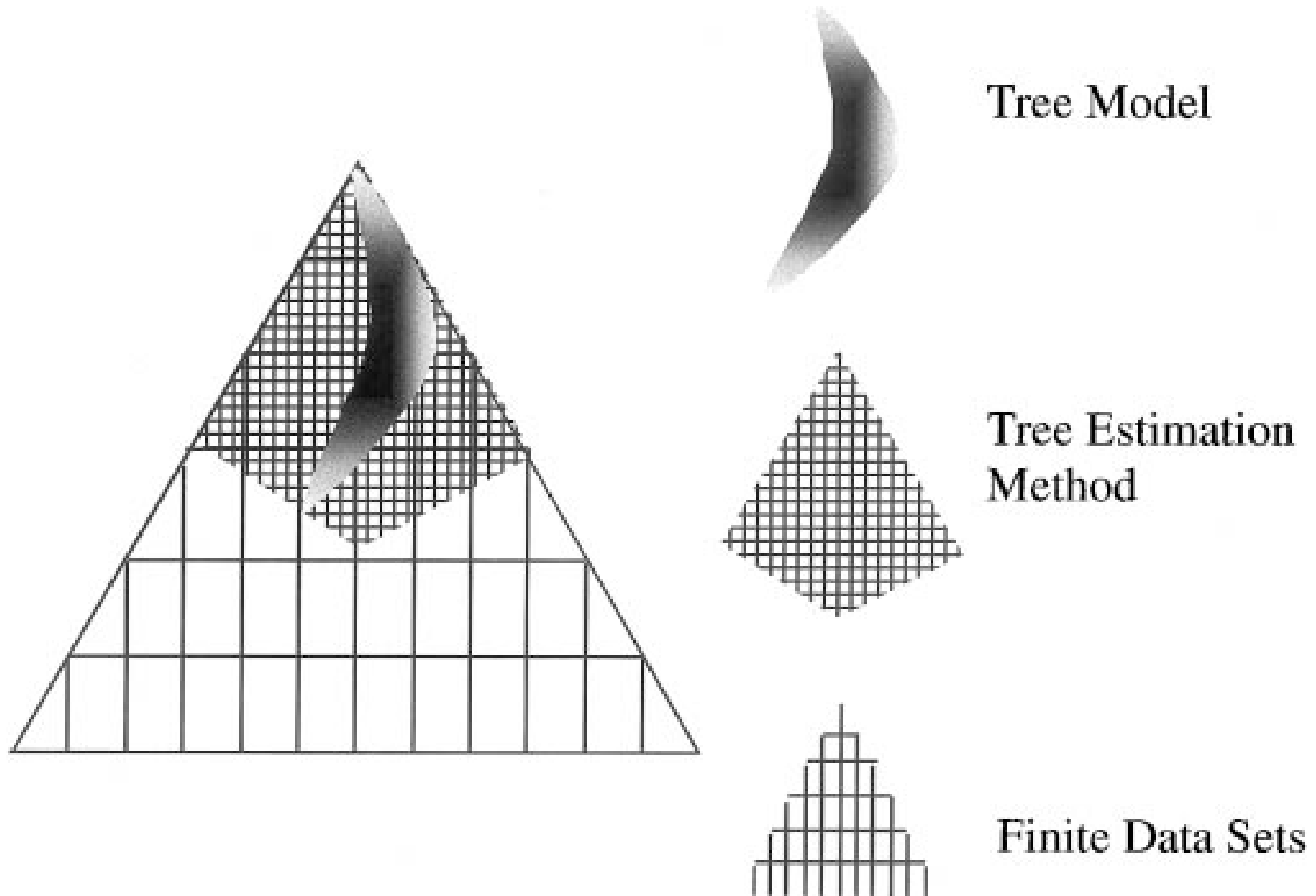
- Null hypotheses:
 - Expected scores are the same → boundary tests. **Non-parametric tests**
 - A tree consistent with the null is correct → tests that use the full info of the model. **Parametric tests**
- How to use variance information:
 - Rely on “raw” bootstrap variability,
 - Invoke assumptions 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

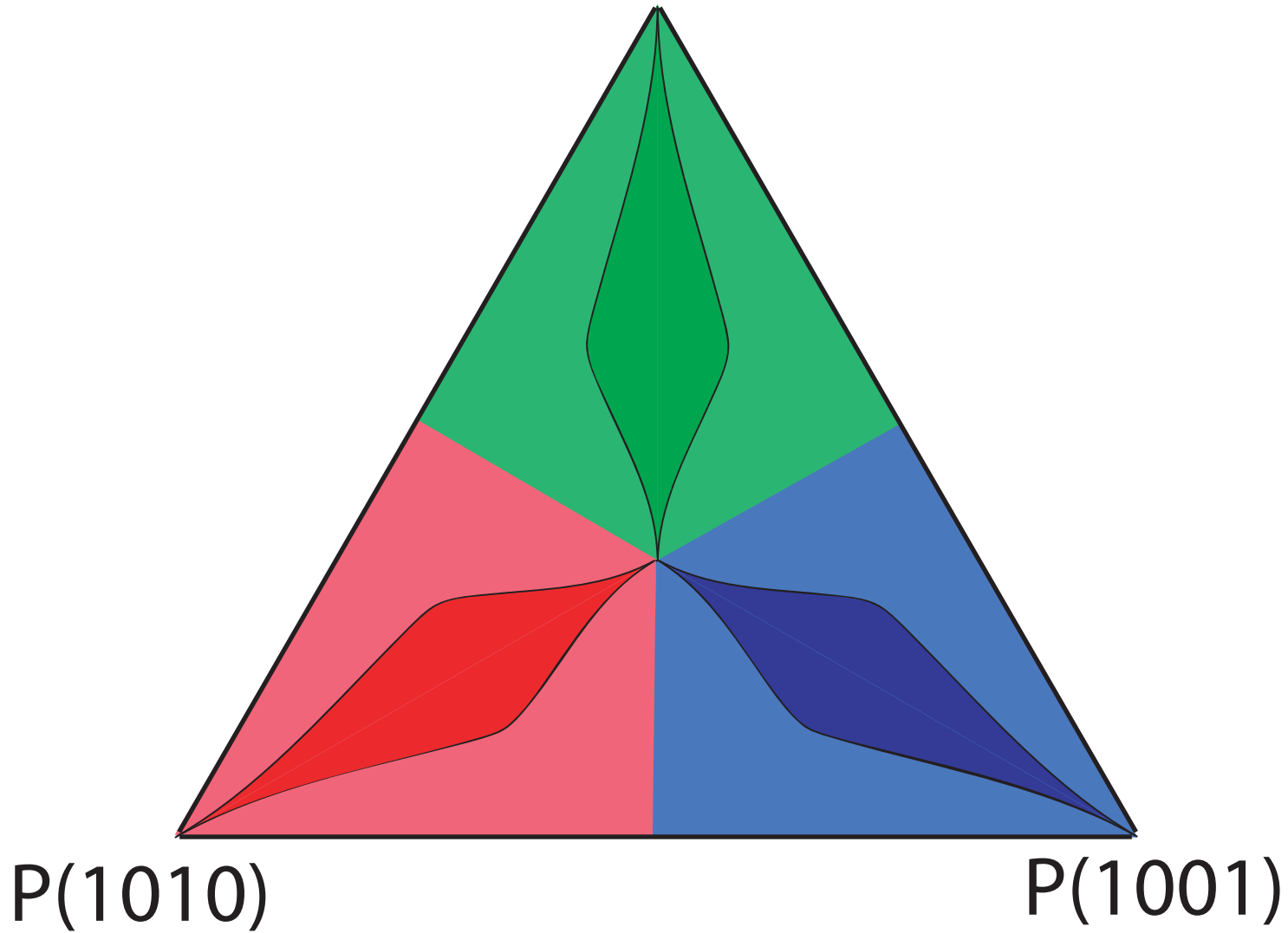
When you use a parametric test, you will usually gain power. But non-parametric tests are more robust to model violation.

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



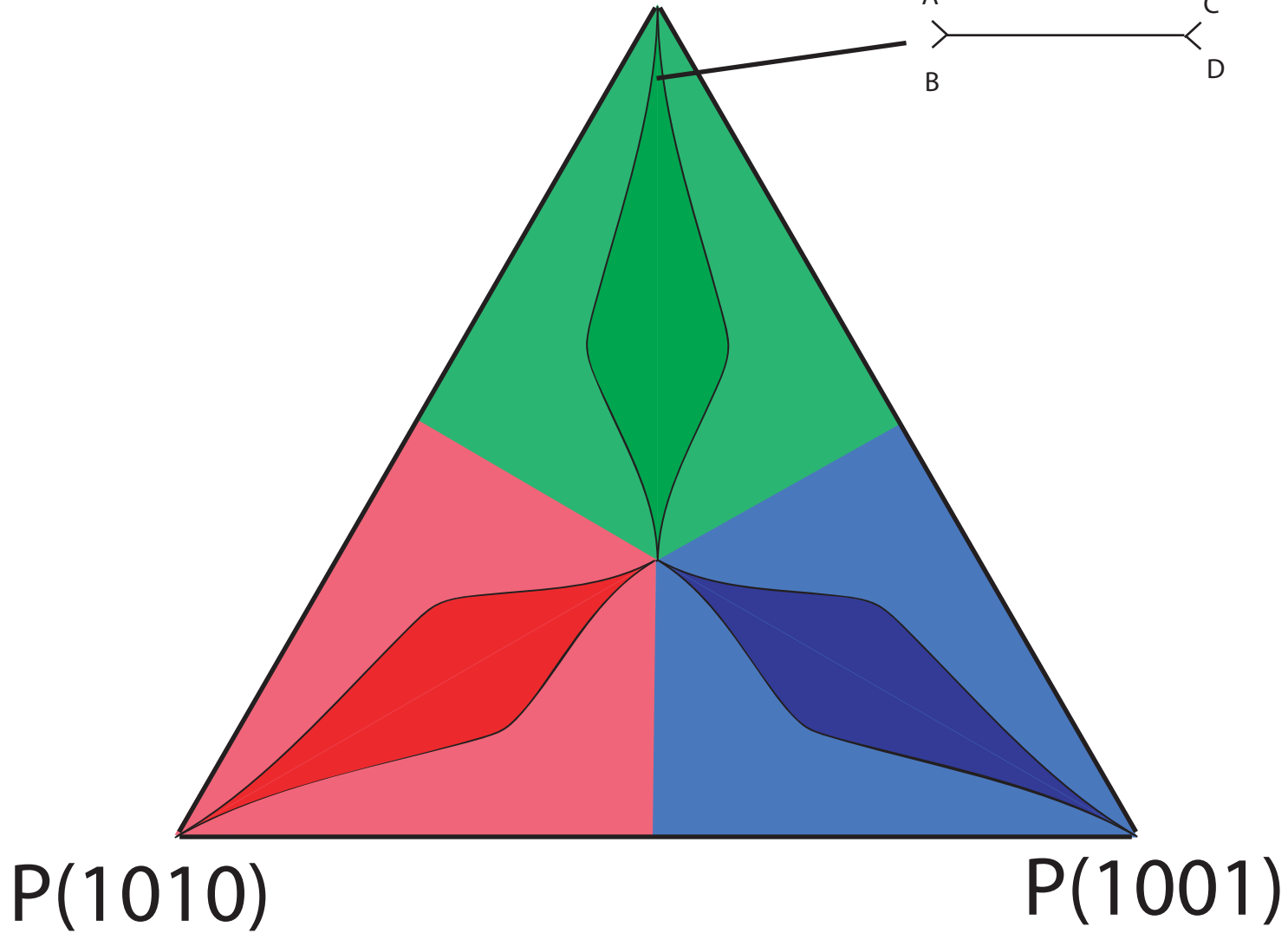
Parsimony-informative Pattern Frequency Space

P(1100)

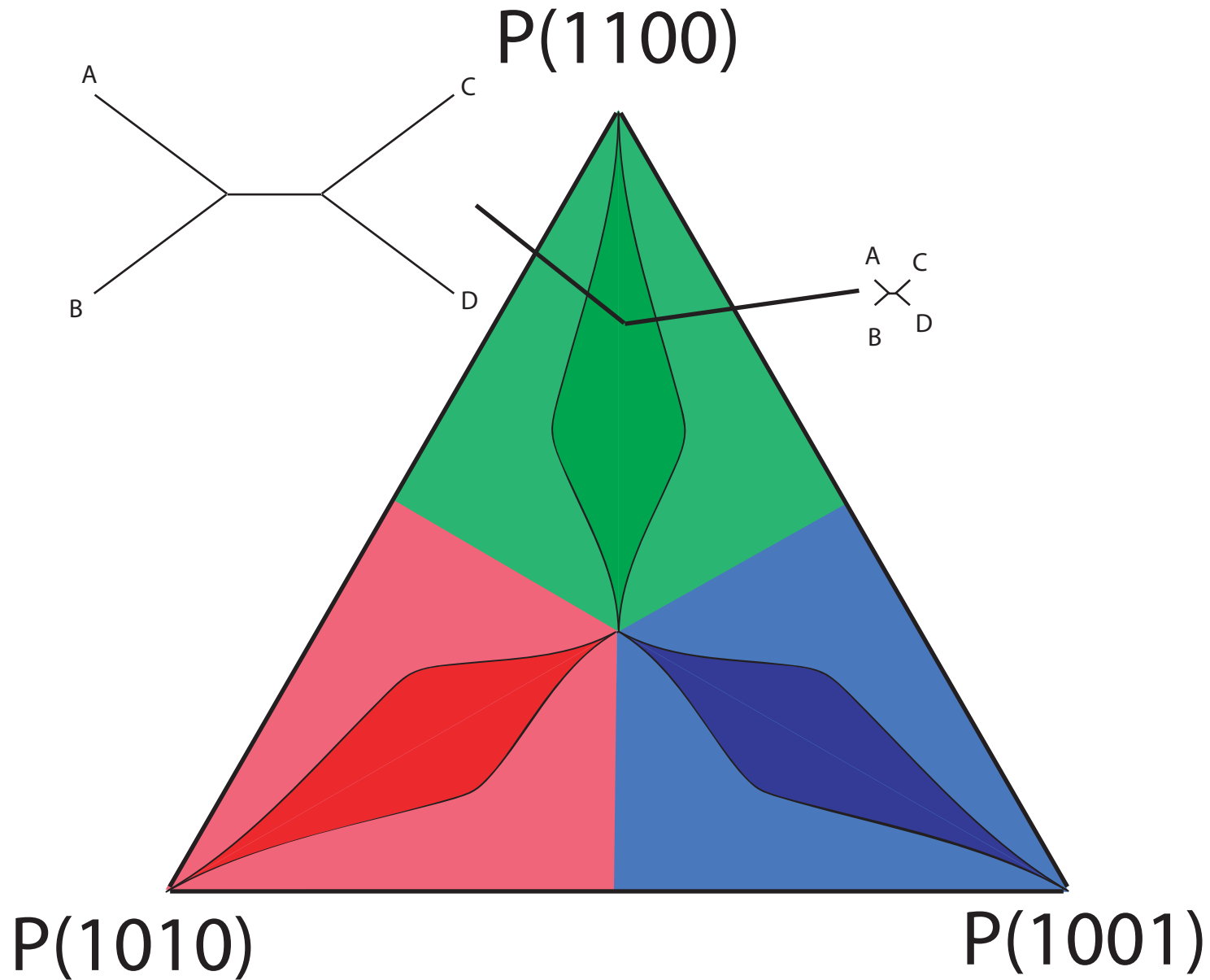


Parsimony-informative Pattern Frequency Space

P(1100)

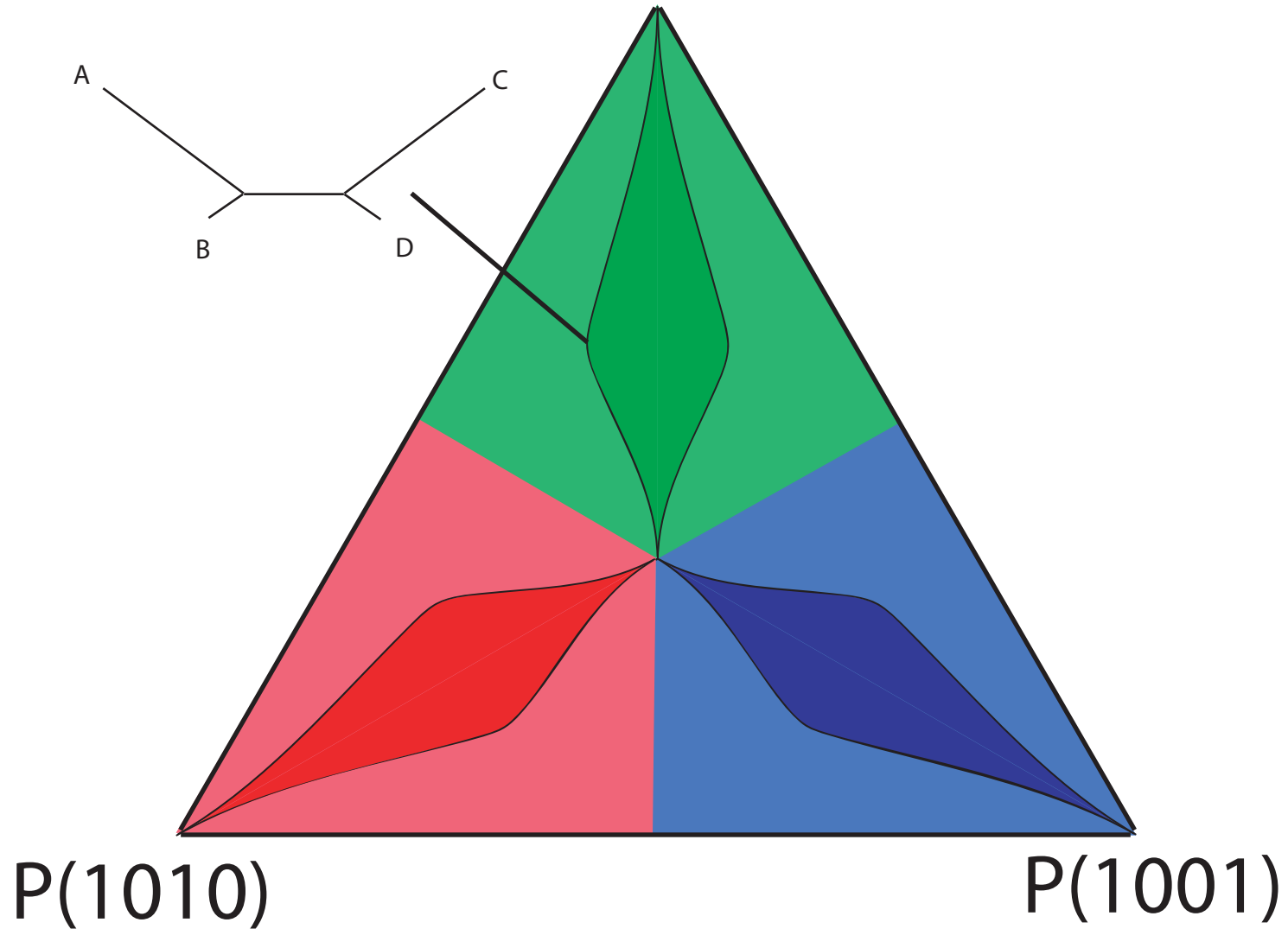


Parsimony-informative Pattern Frequency Space

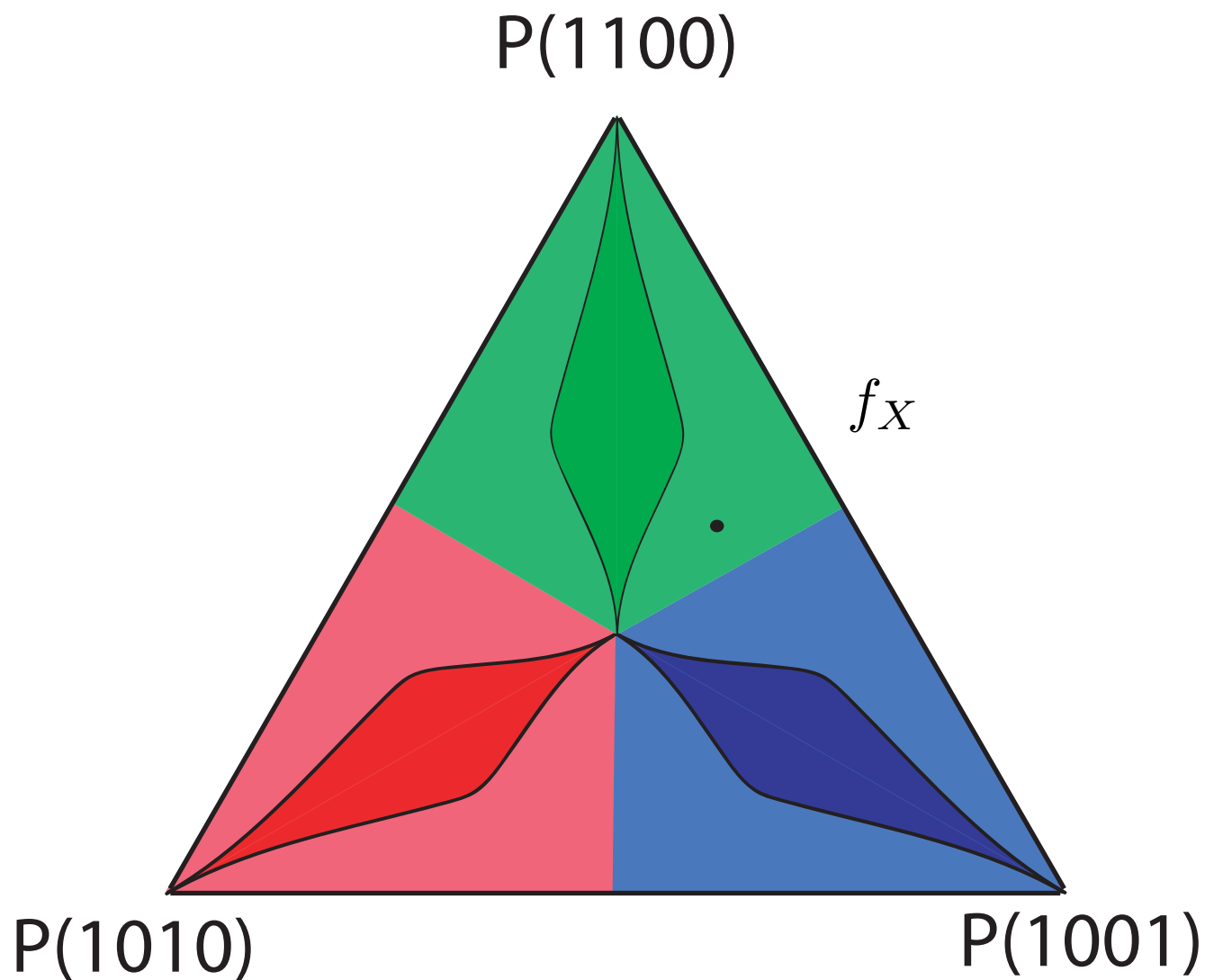


Parsimony-informative Pattern Frequency Space

P(1100)



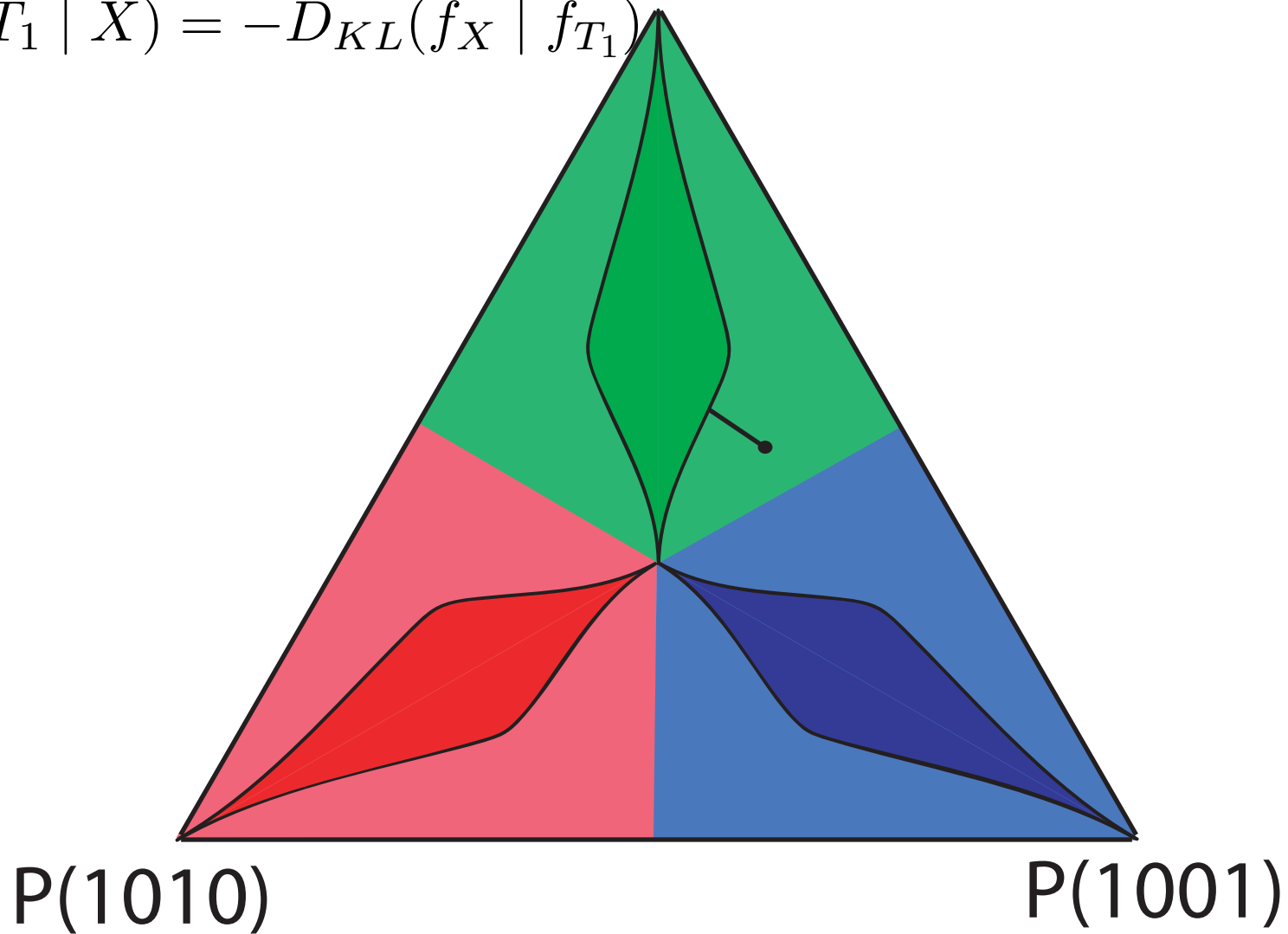
Pattern Frequency Space With Observed Data



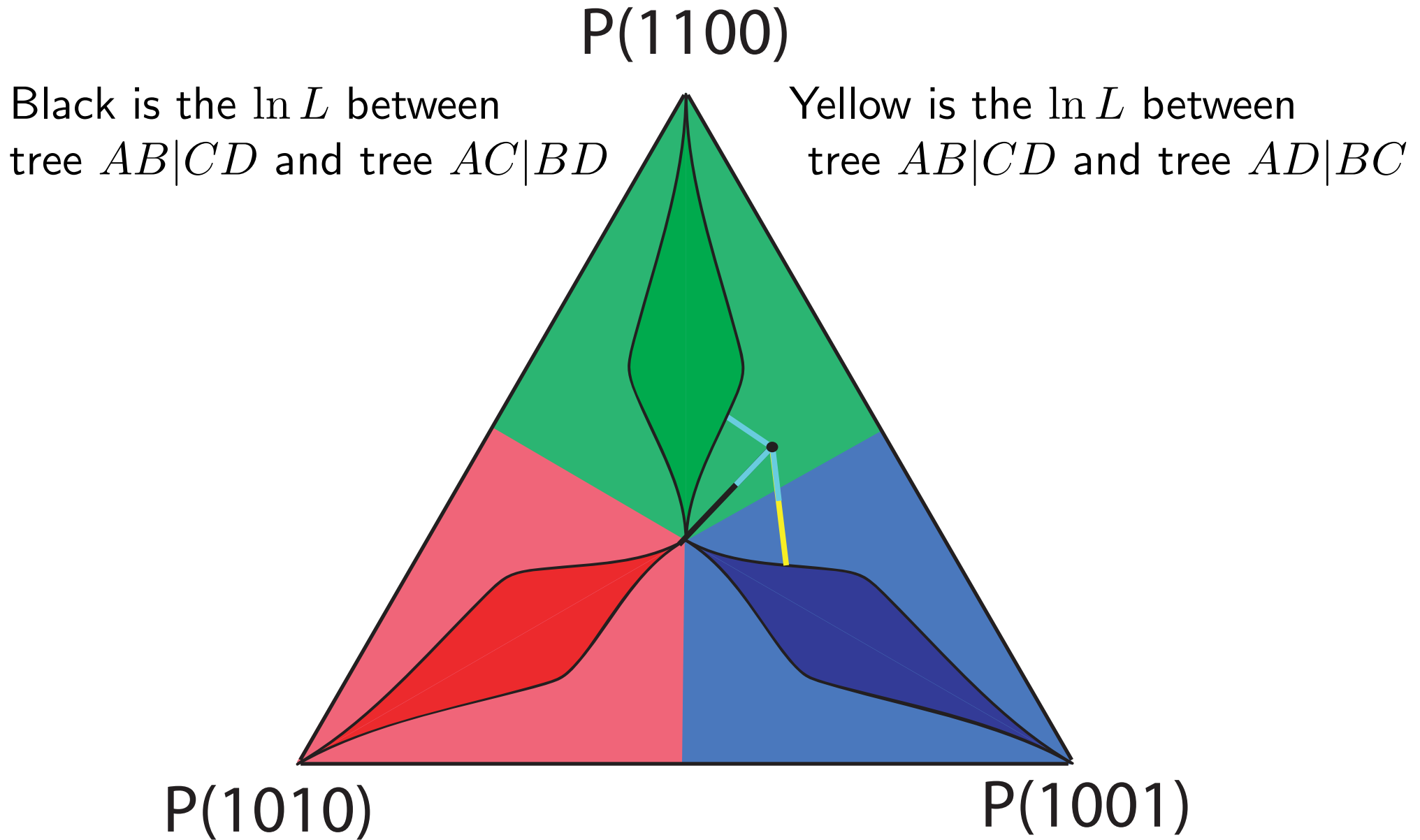
ML scores in Pattern Frequency Space

P(1100)

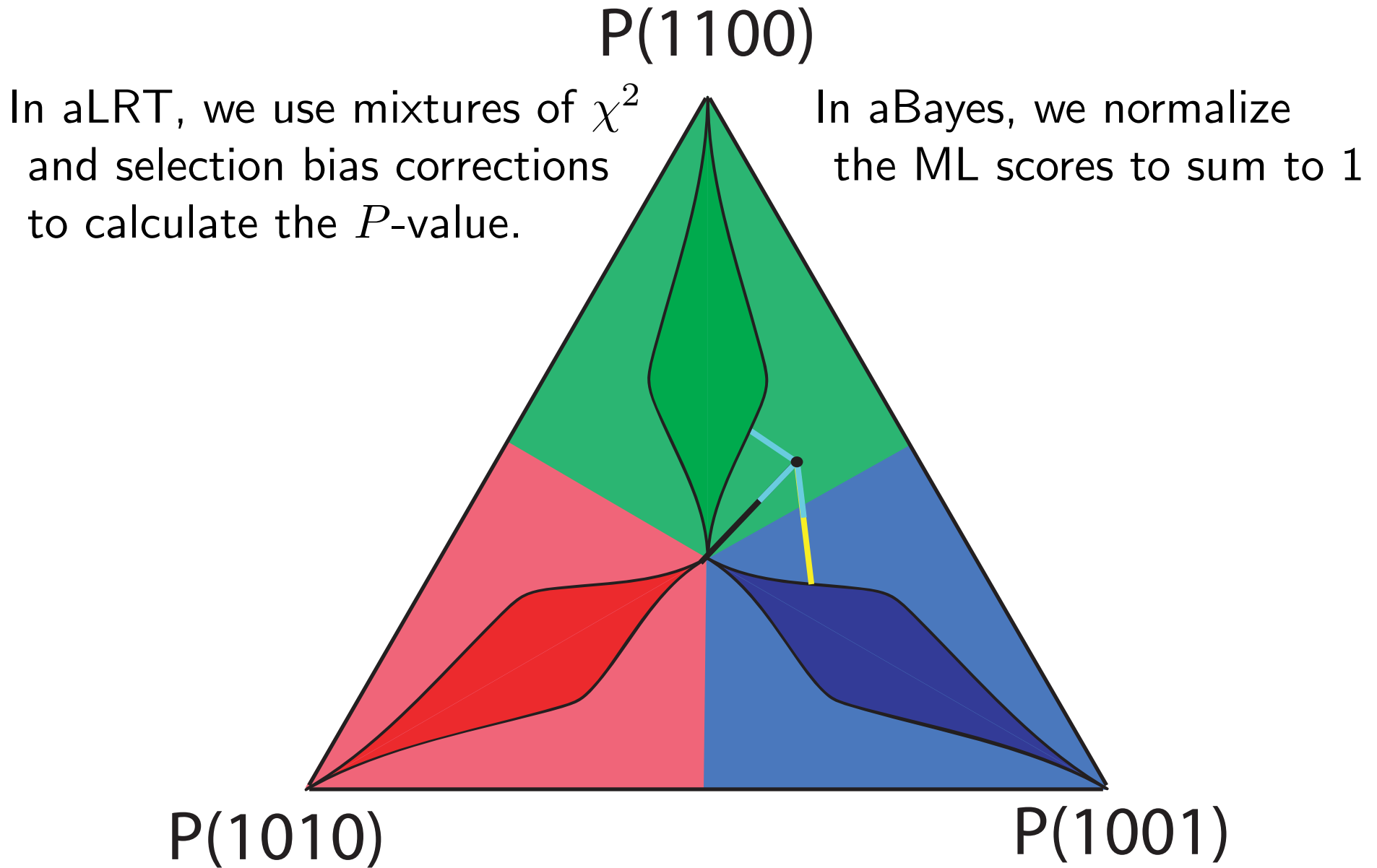
$$\ln L(T_1 | X) = -D_{KL}(f_X | f_{T_1})$$



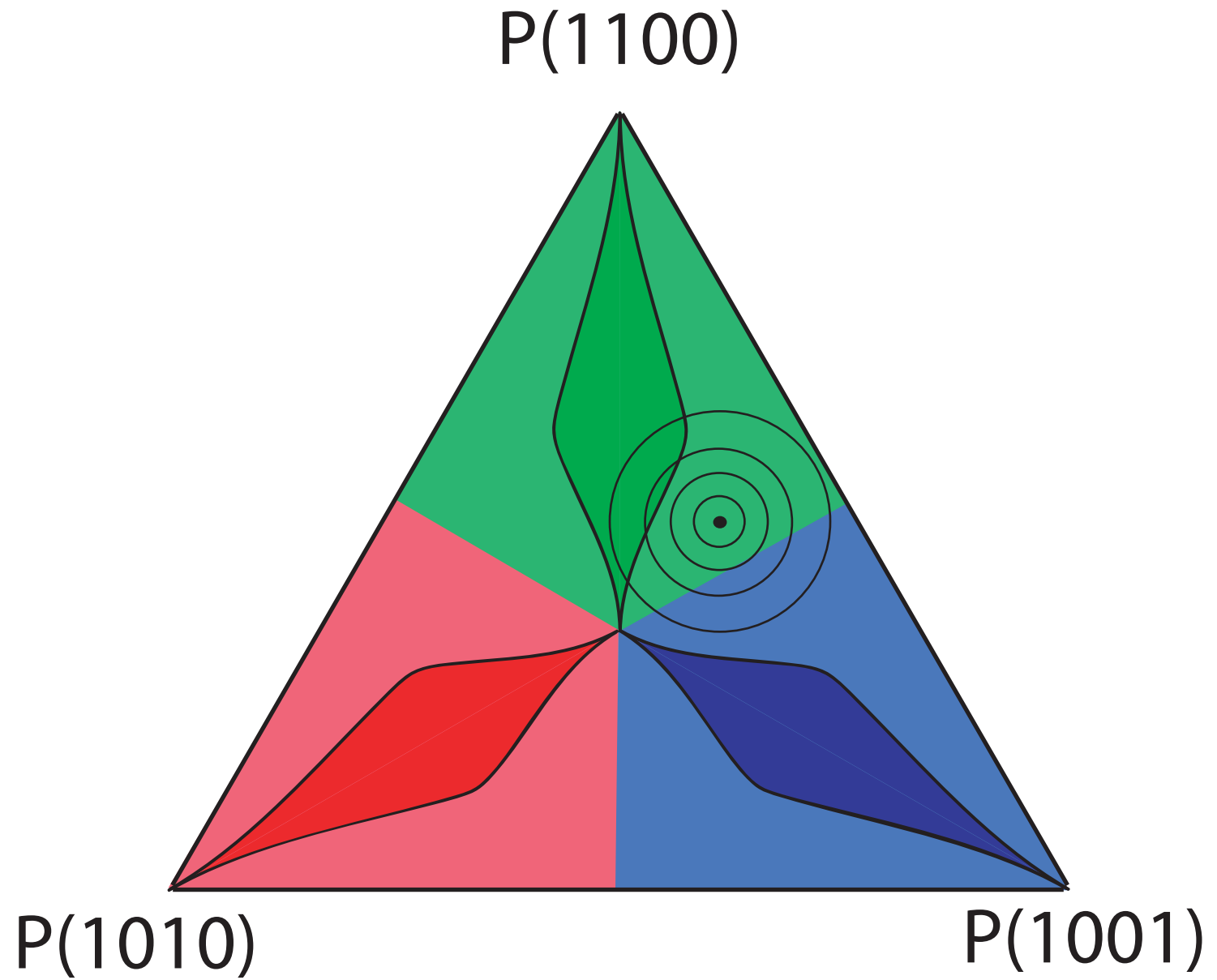
LR statistics in Pattern Frequency Space



aLRT and aBayes in Pattern Frequency Space

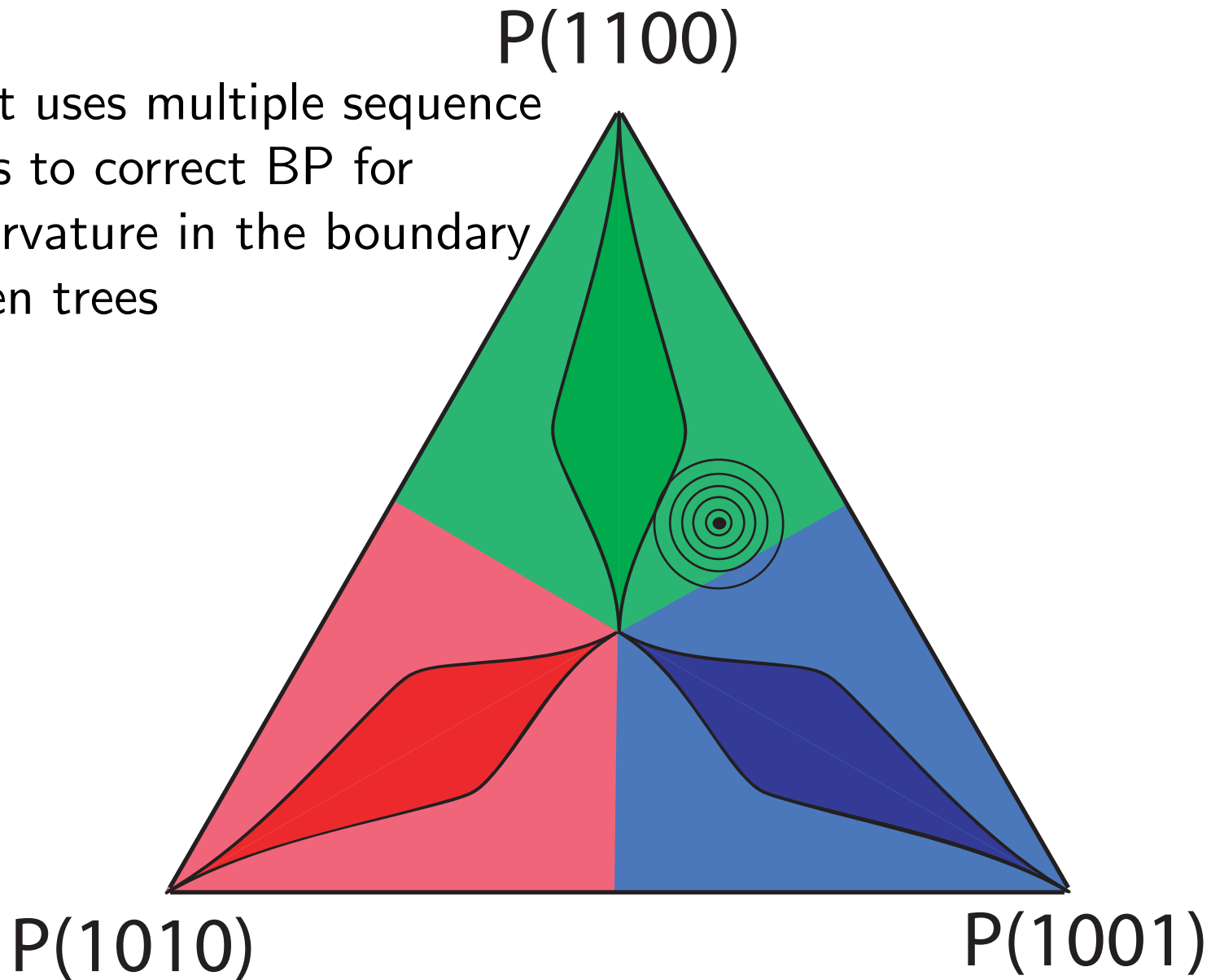


Non-parametric Bootstrapping in Pattern Frequency Space



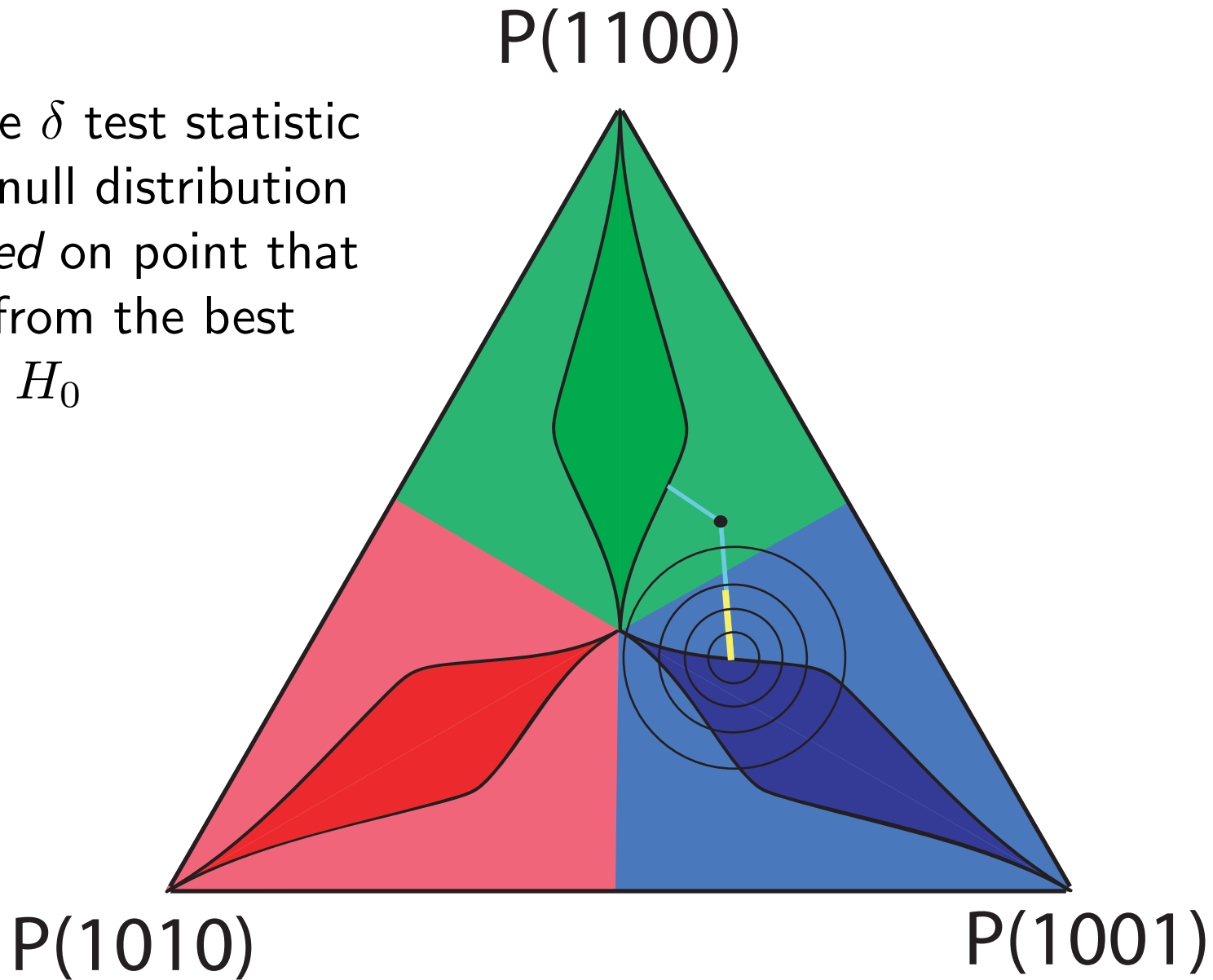
Bootstrapping in Pattern Frequency Space (if you had more data)

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



Parametric bootstrapping in Pattern Frequency Space

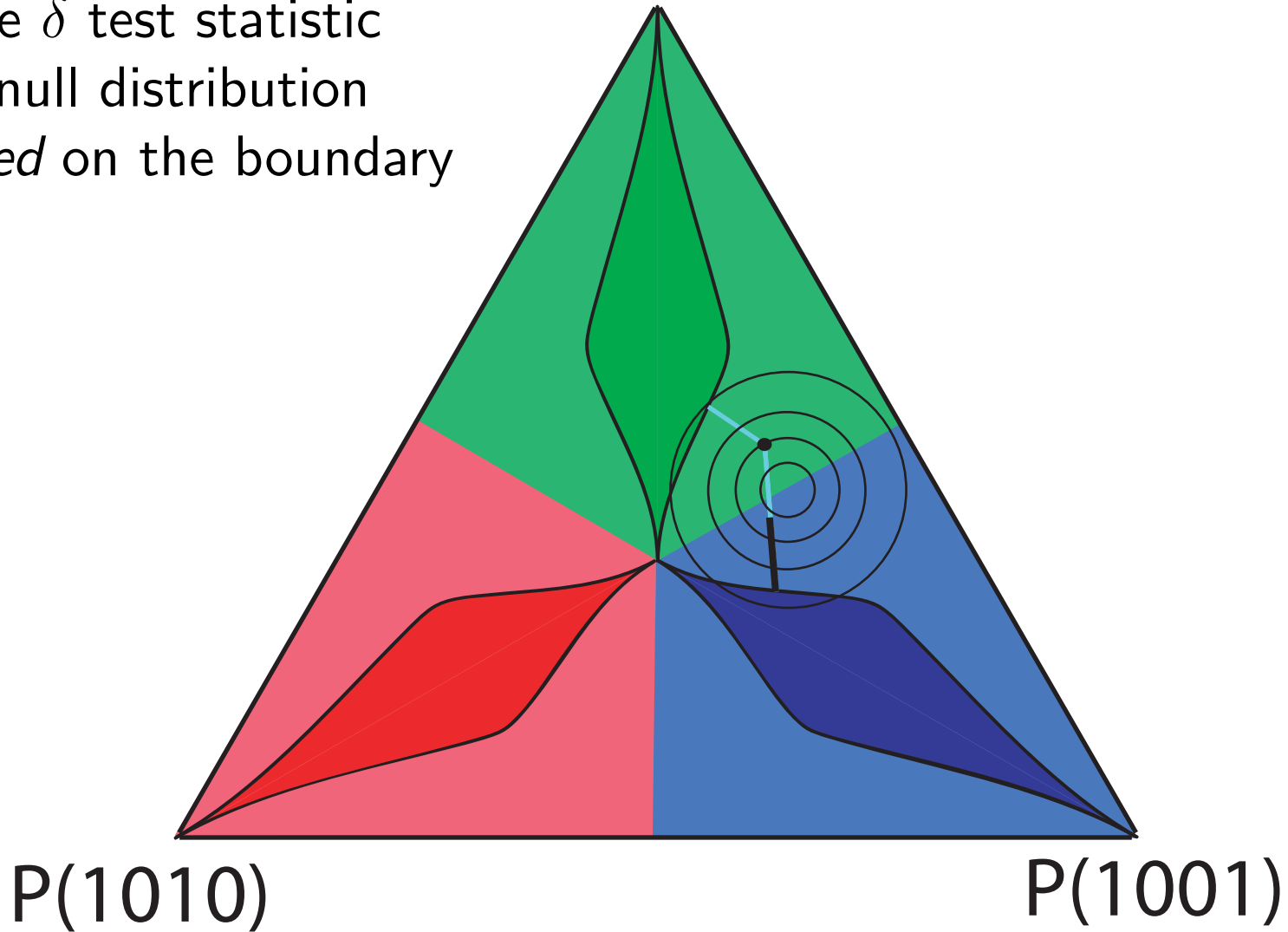
Uses the δ test statistic and a null distribution centered on point that arises from the best tree in H_0



KH Test in Pattern Frequency Space

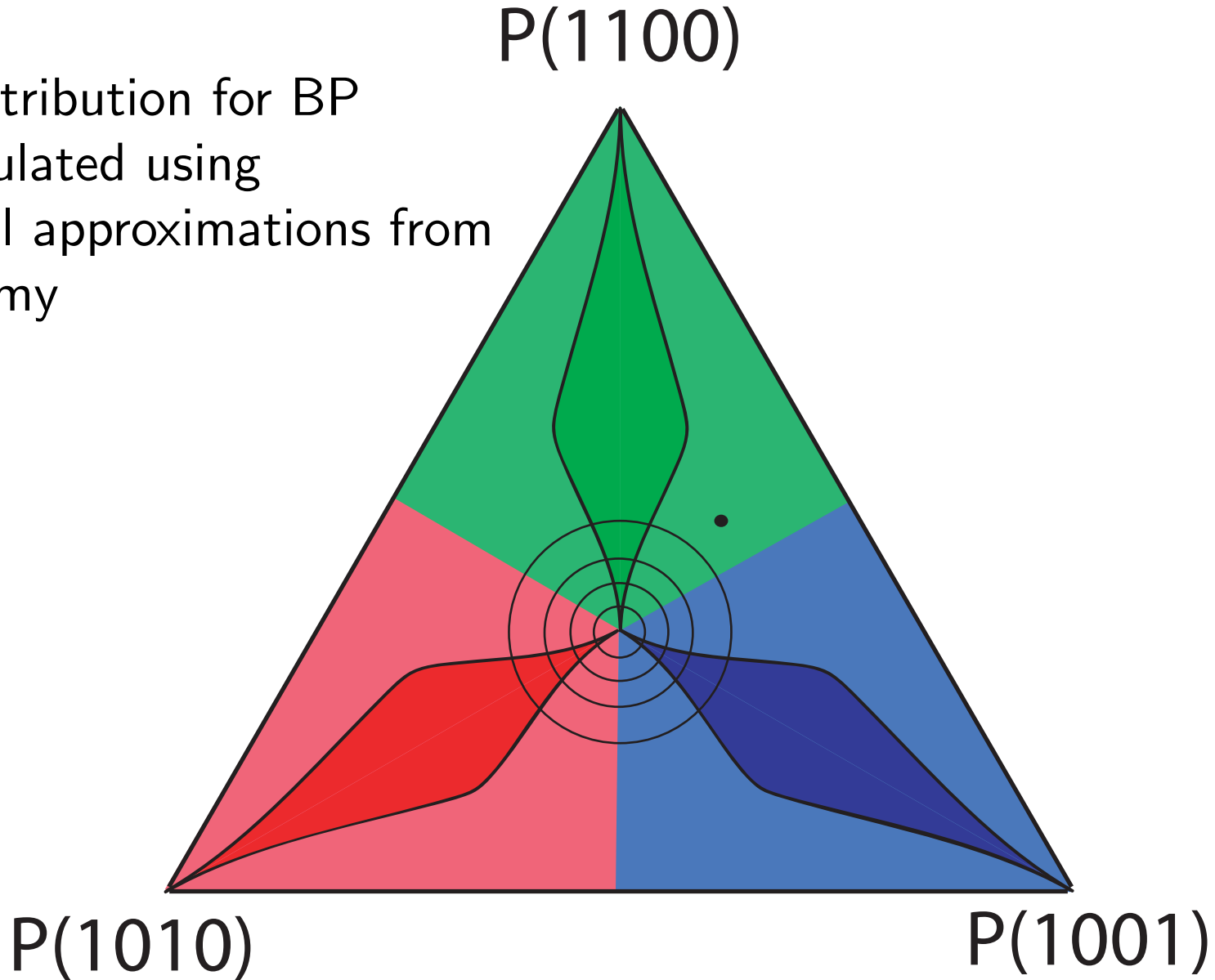
P(1100)

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



aBP in Pattern Frequency Space

Null distribution for BP
is calculated using
Normal approximations from
polytomy



Summary - Part 2

A (very) wide variety of tests differ by:

- Null hypotheses:
 - Expected scores are the same → boundary tests. **Non-parametric tests**
 - A tree consistent with the null is correct → 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.

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.

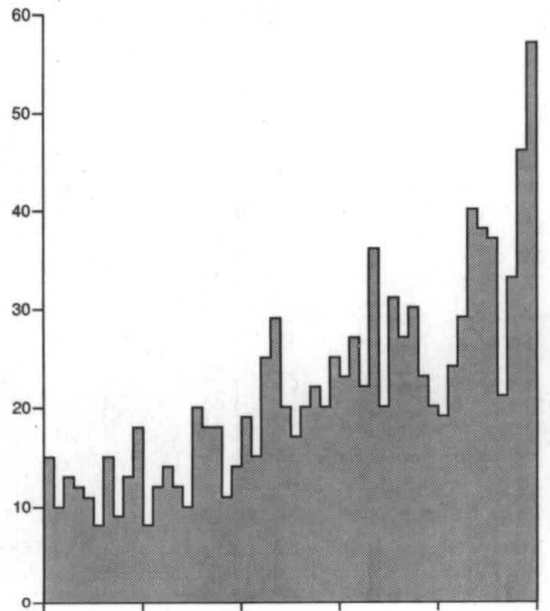
Suzuki, Y., Glazko, G. V., and Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *PNAS*, 99:16138–16143.

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.

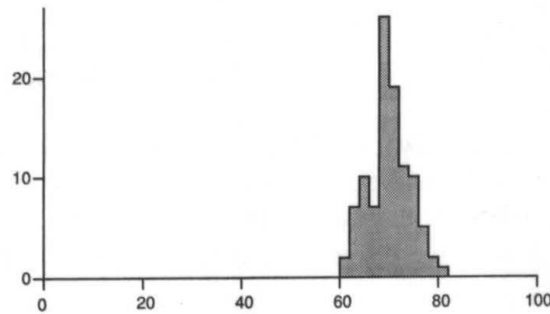
Wilcox, T. P., Zwickl, D. J., Heath, T., and Hillis, D. M. (2002). Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Molecular Phylogenetics and Evolution*, 25:361–371.

Bootstrapping as a noisy measure of repeatability

bootstrap
values from
many simulations



repeated
simulation



% recovering tree

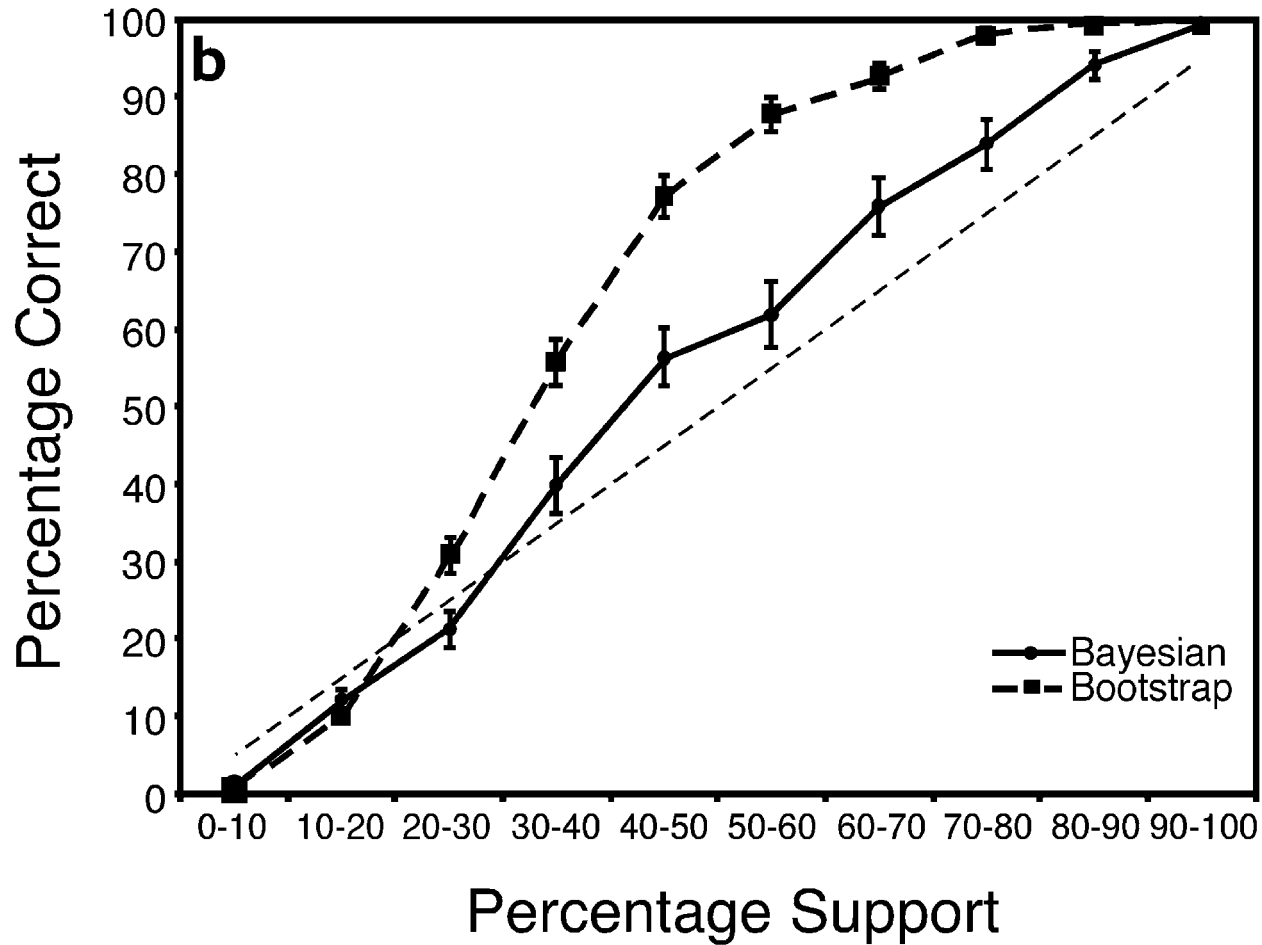
Simulation study of
Hillis and Bull (1993)

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 probabilities 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

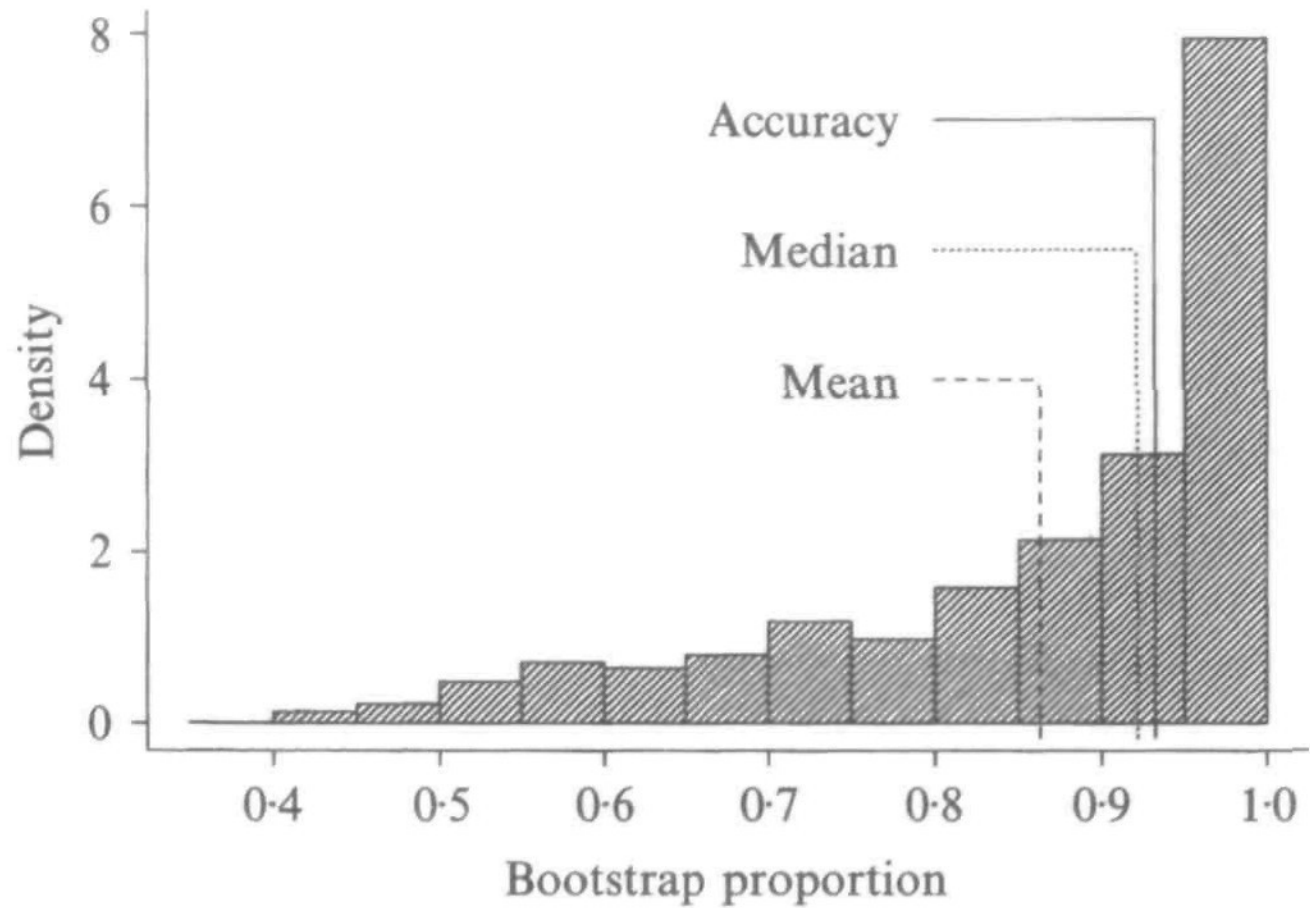
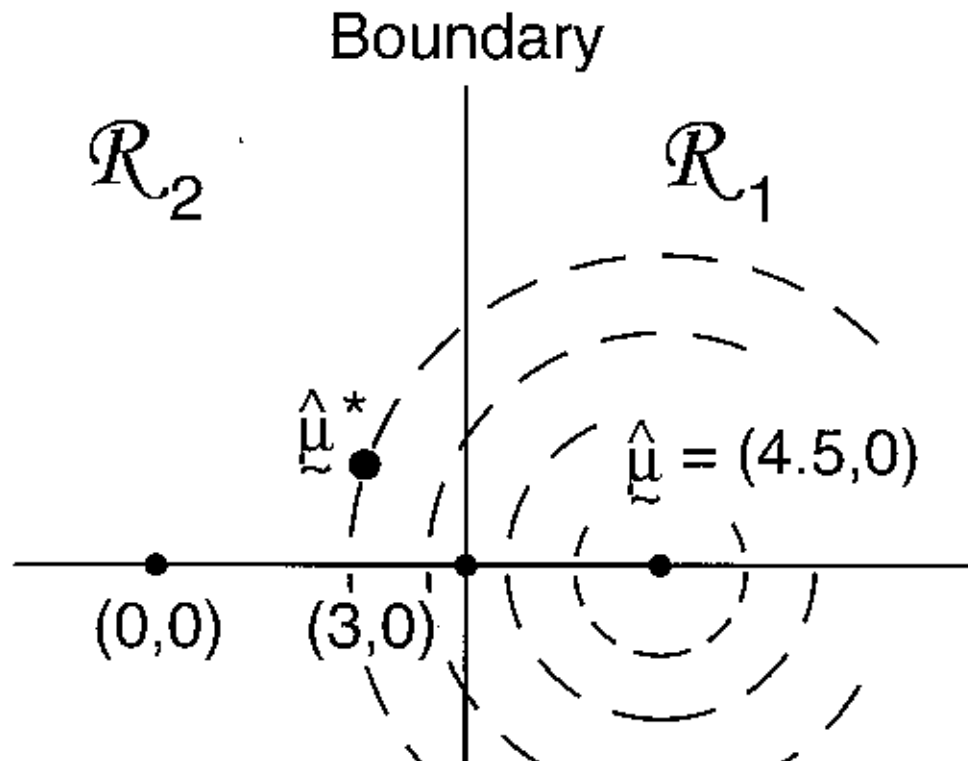


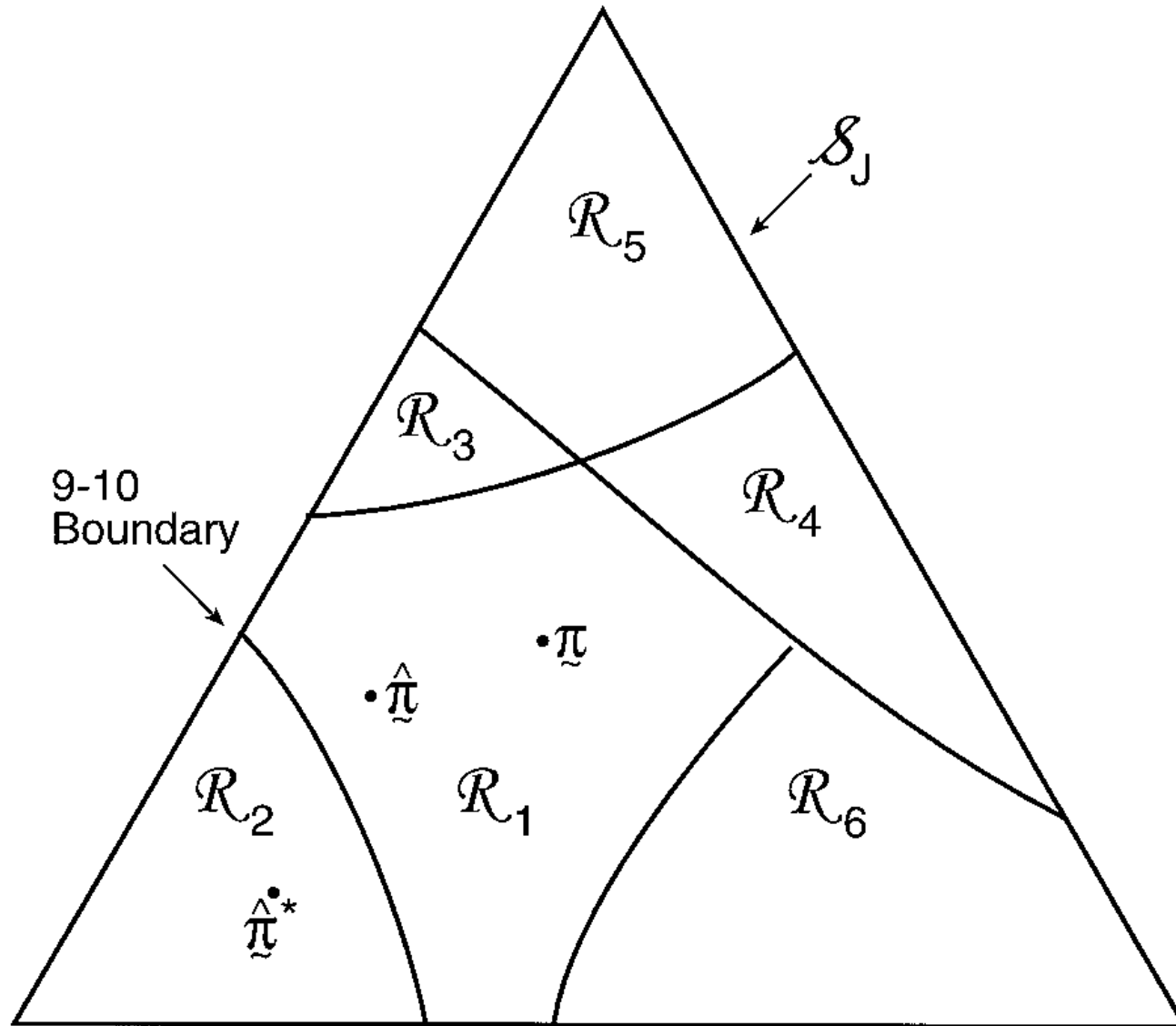
Figure 4 from Newton (1996)

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 \mathcal{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 $\mu \in \mathcal{R}_1$, if we begin with an “uninformative” prior density for μ .



Efron et al. (1996) view of tree space

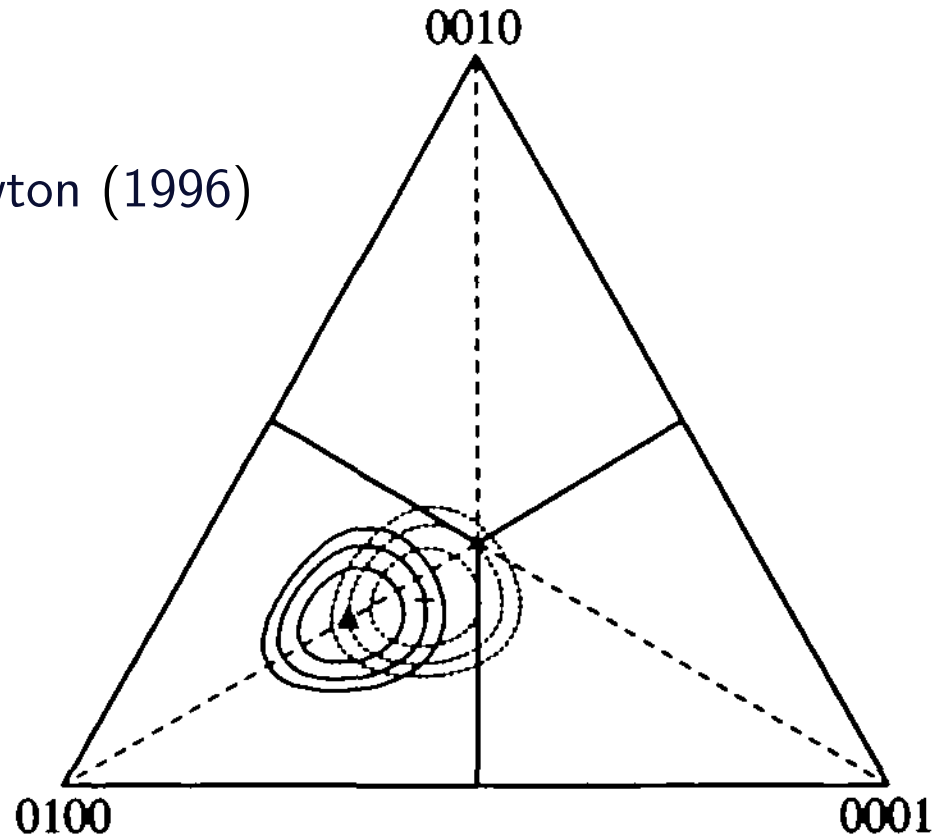


What did Efron et al. (1996) say (and mean)?

- 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)

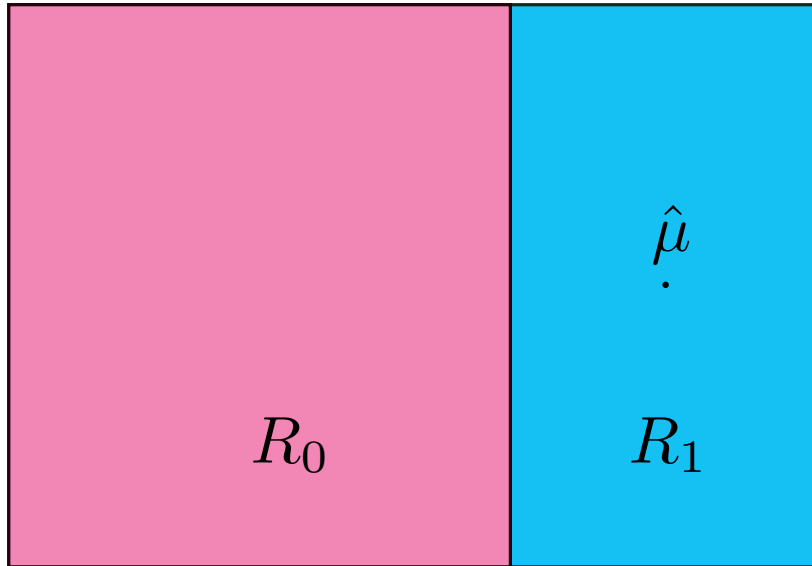


Darker ovals indicate probability contours for datasets given the truth
(note that repeatability $\approx 100\%$)

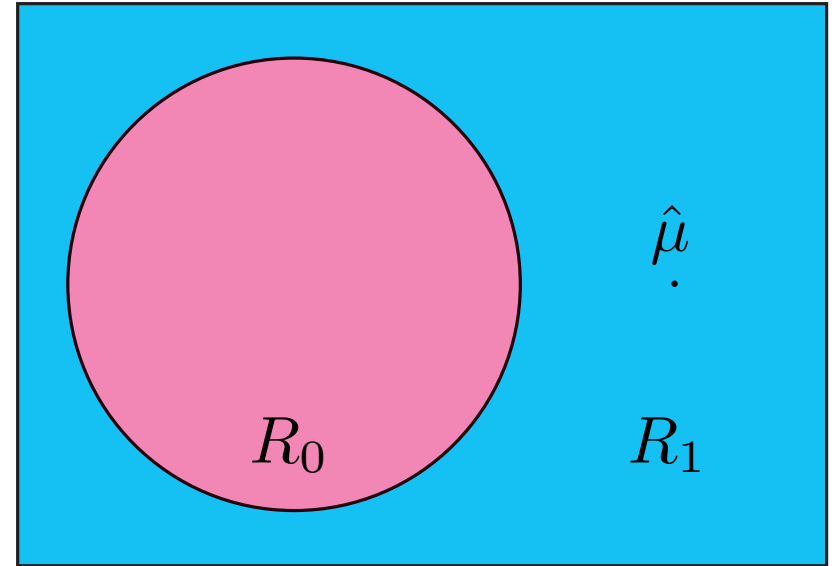
Lighter ovals show probability contours for bootstrapping for one dataset.

Many real datasets will have BP much $< 100\%$

Case 1

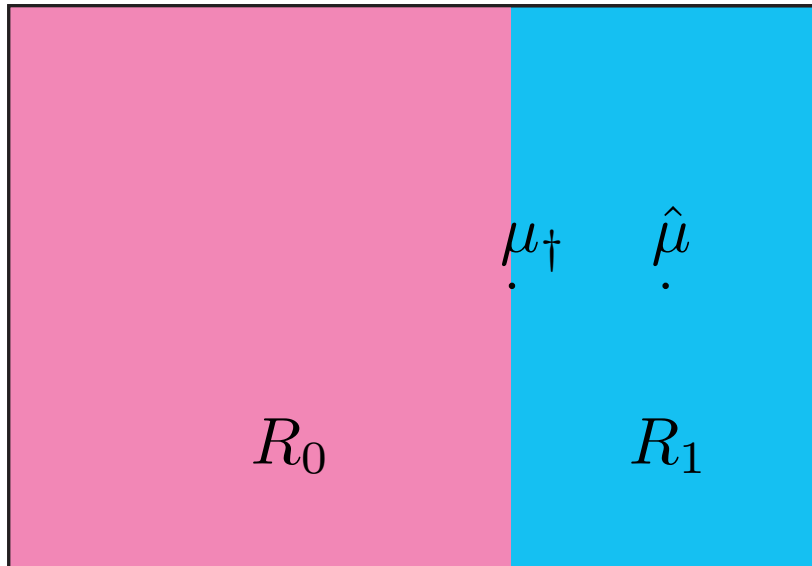


Case 2

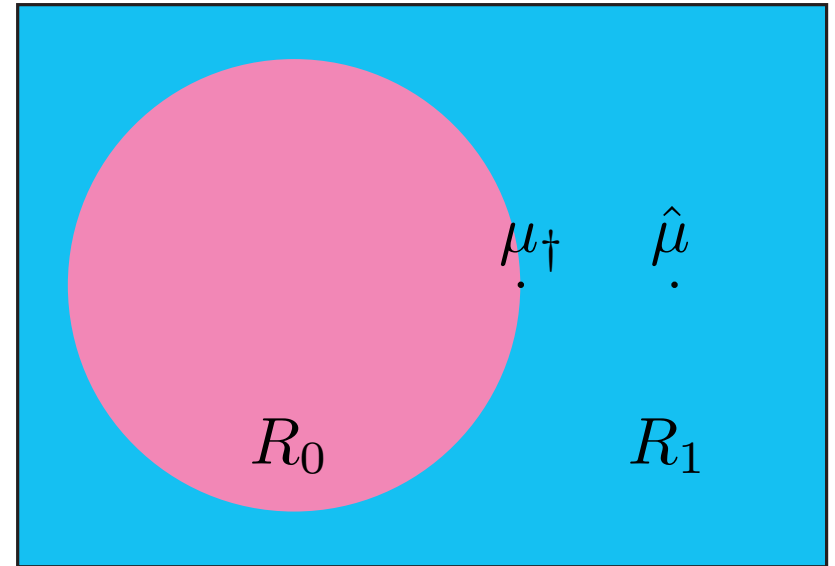


$\hat{\mu}$ is the best point calculated from the data

Case 1

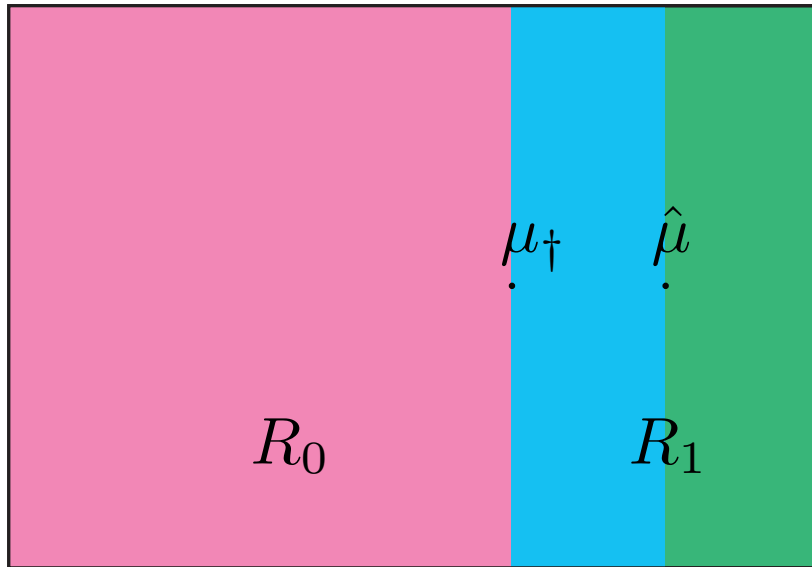


Case 2

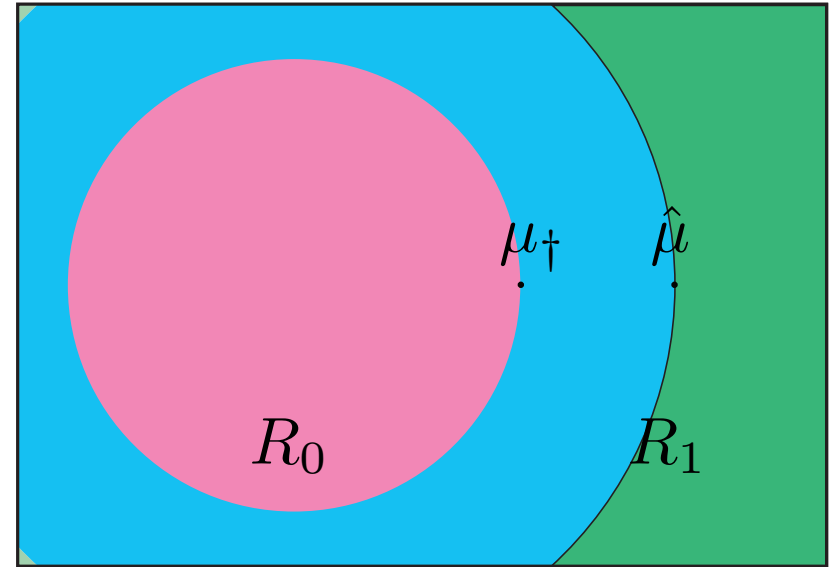


$\hat{\mu}$ is the best point calculated from the data
 μ_{\dagger} is least-favorable condition (LFC) point in R_0

Case 1

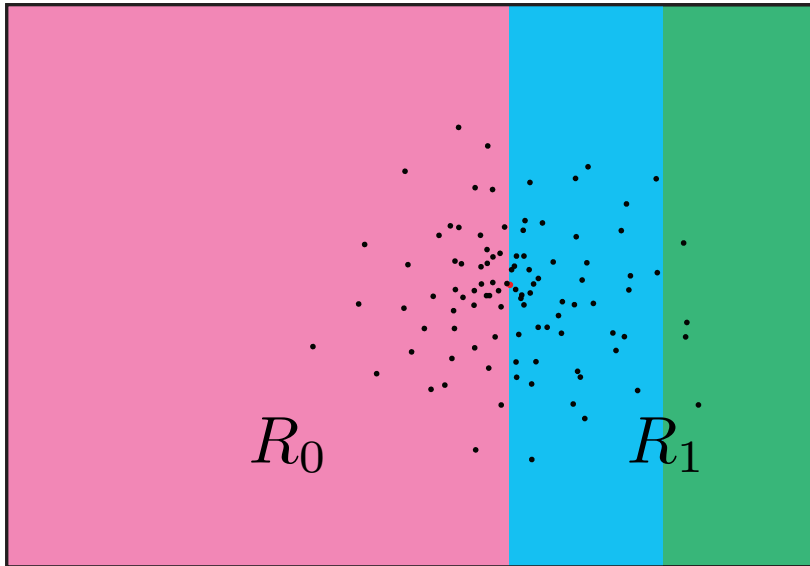


Case 2

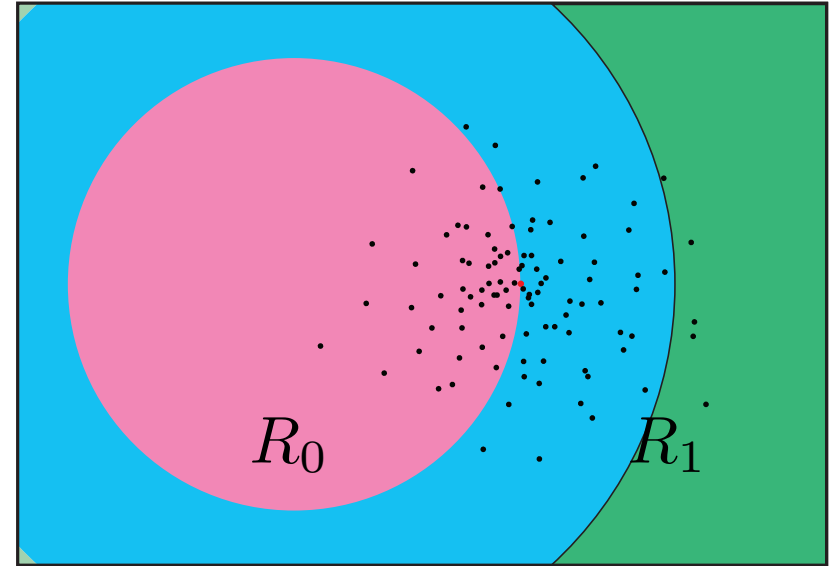


$\hat{\mu}$ is the best point calculated from the data
 μ_{\dagger} is least-favorable condition (LFC) point in R_0
 green areas are the tails - they correspond to values of the test
 statistic more extreme than $\hat{\mu}$ (relative to that $\mu \in R_0$)

Case 1

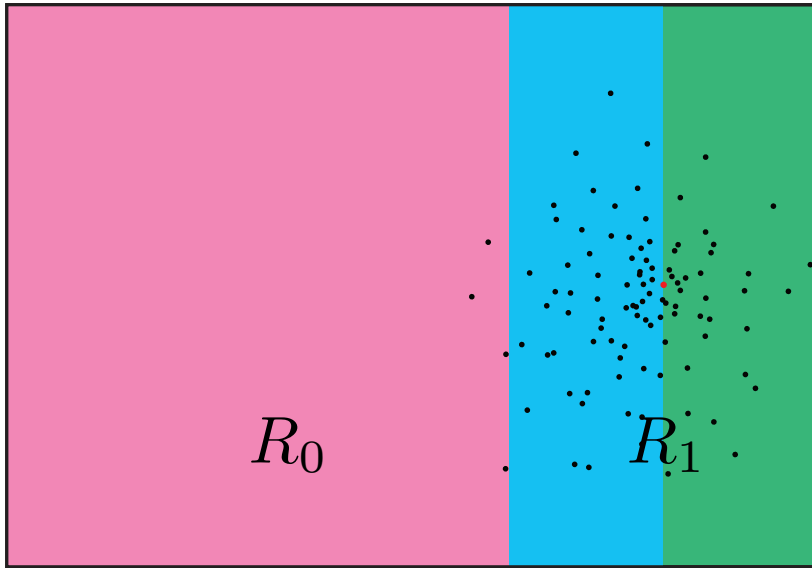


Case 2

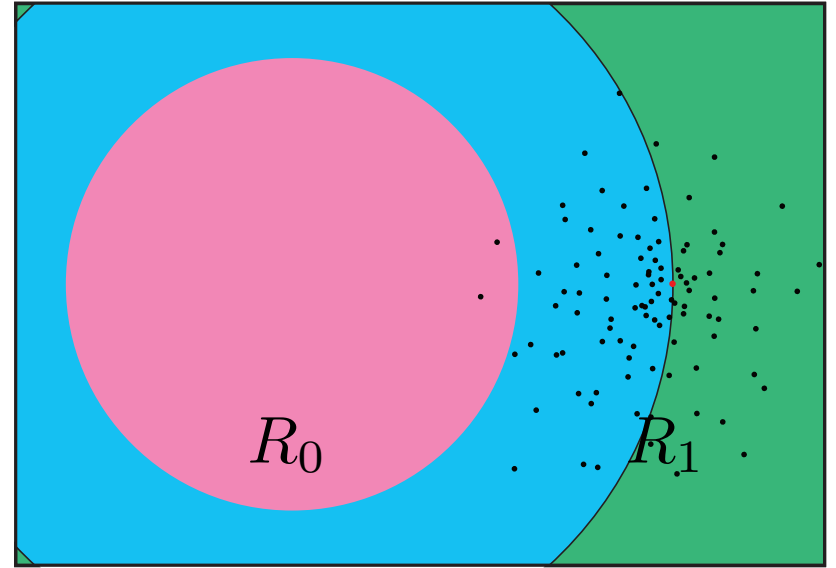


$\hat{\mu}$ is the best point calculated from the data
 μ_{\dagger} is least-favorable condition (LFC) point in R_0
Case 1 P-value < the P-value in Case 2

Case 1

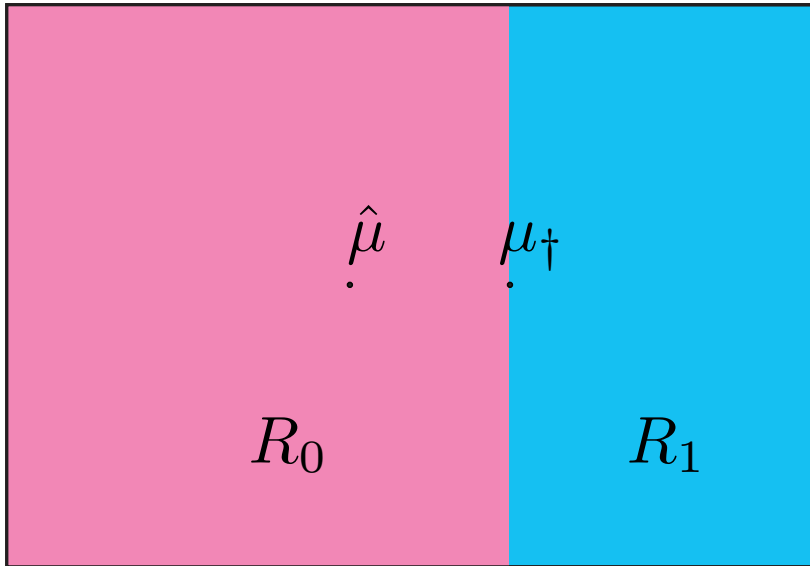


Case 2

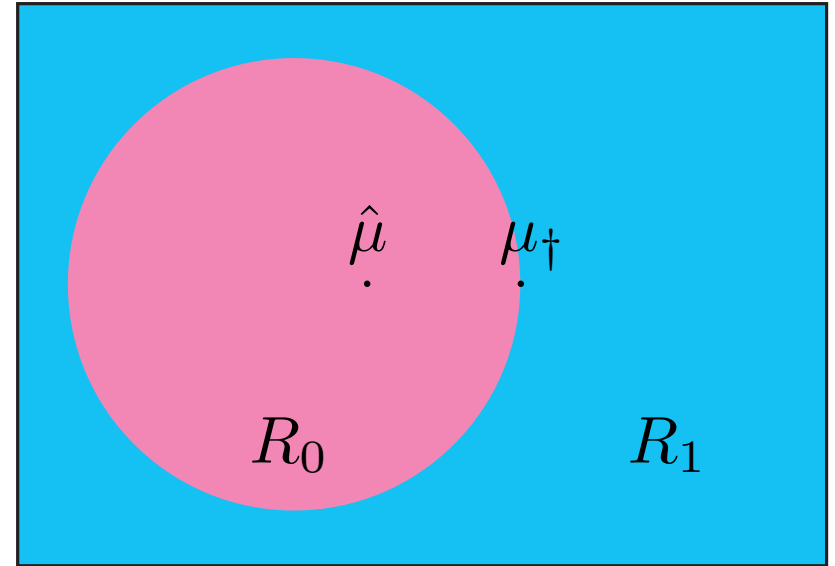


In case 1 - the bootstrap proportion is a good estimate of the P-value
In case 2 - the bootstrap proportion underestimates the P-value

Case 3

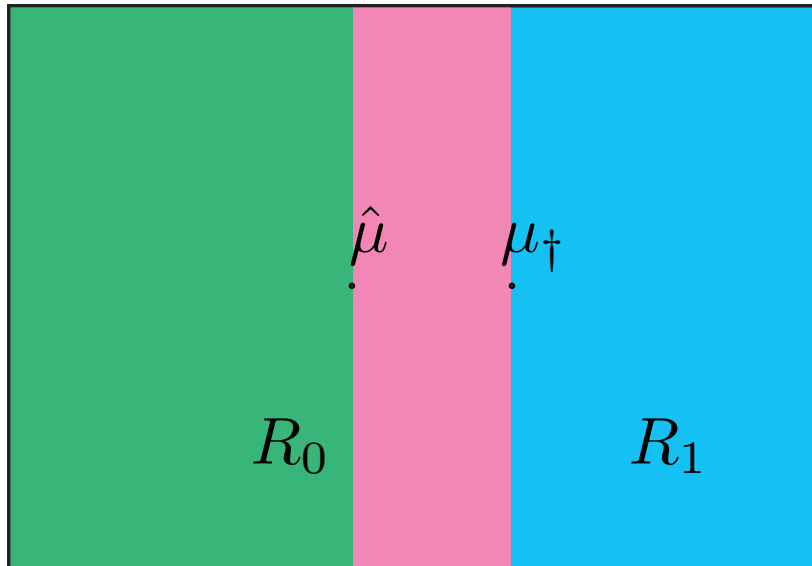


Case 4

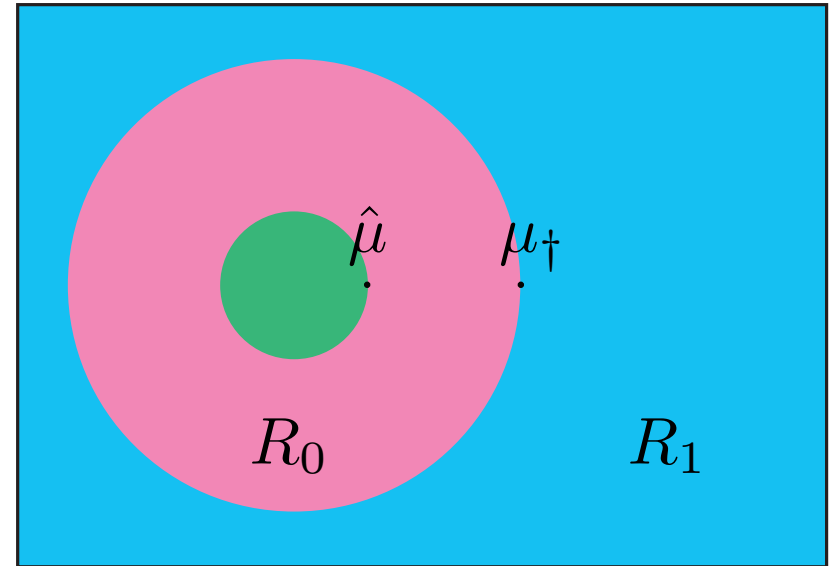


$\hat{\mu}$ is the best point calculated from the data
 μ_{\dagger} is least-favorable condition (LFC) point in R_1

Case 3

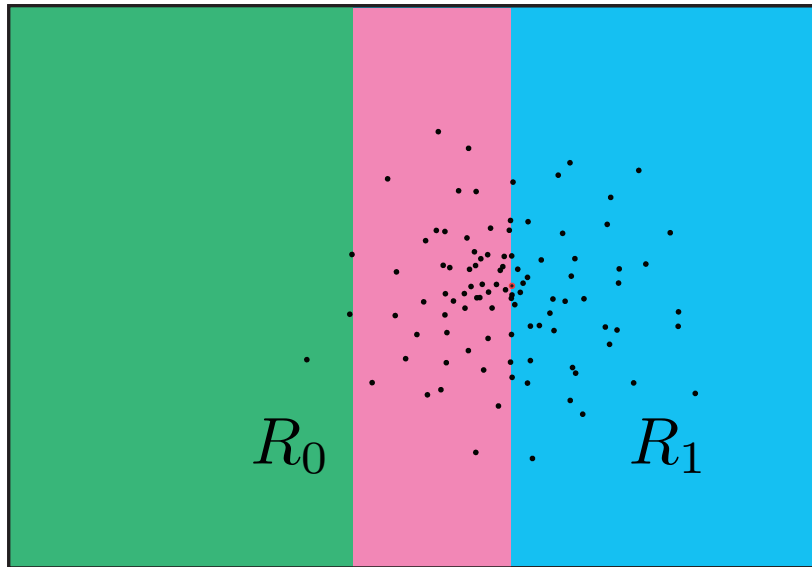


Case 4

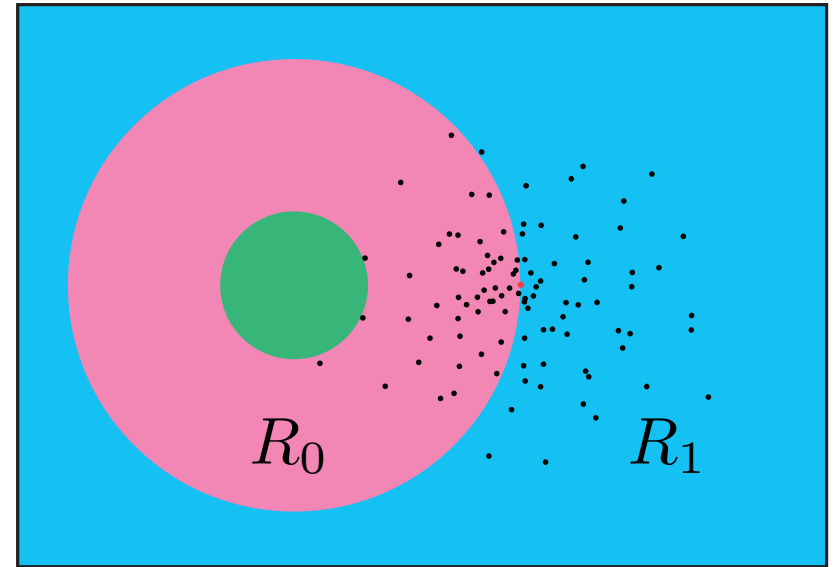


$\hat{\mu}$ is the best point calculated from the data
 μ_{\dagger} is least-favorable condition (LFC) point in R_0
 green areas are the tails - they correspond to values of the test
 statistic more extreme than $\hat{\mu}$ (relative to that $\mu \in R_1$)

Case 3

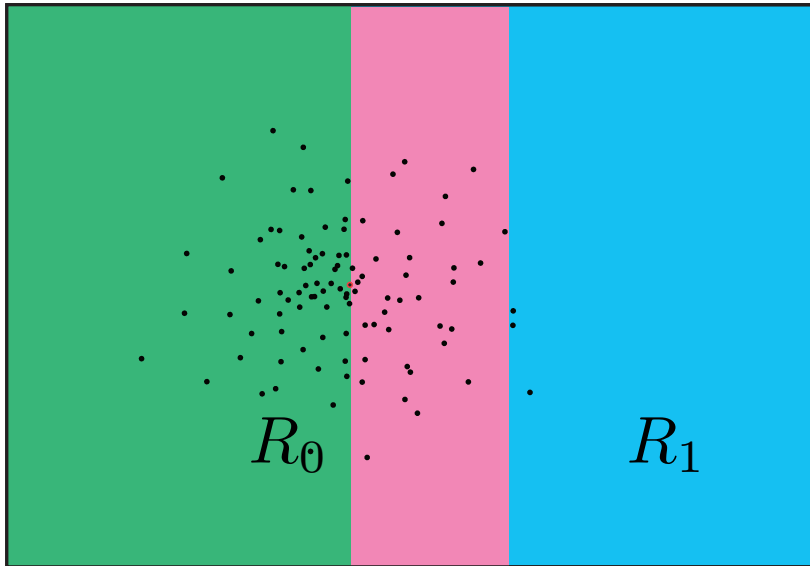


Case 4

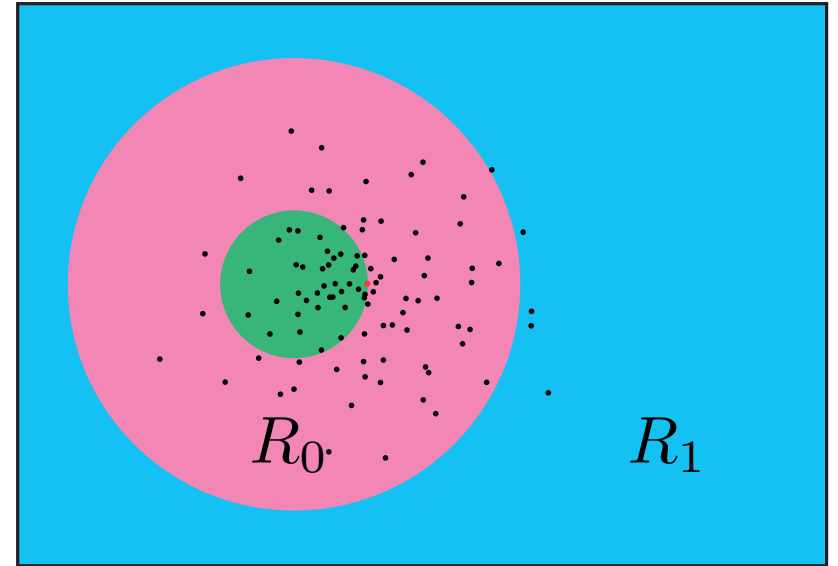


Case 3 P-value $>$ the P-value in Case 4

Case 3



Case 4

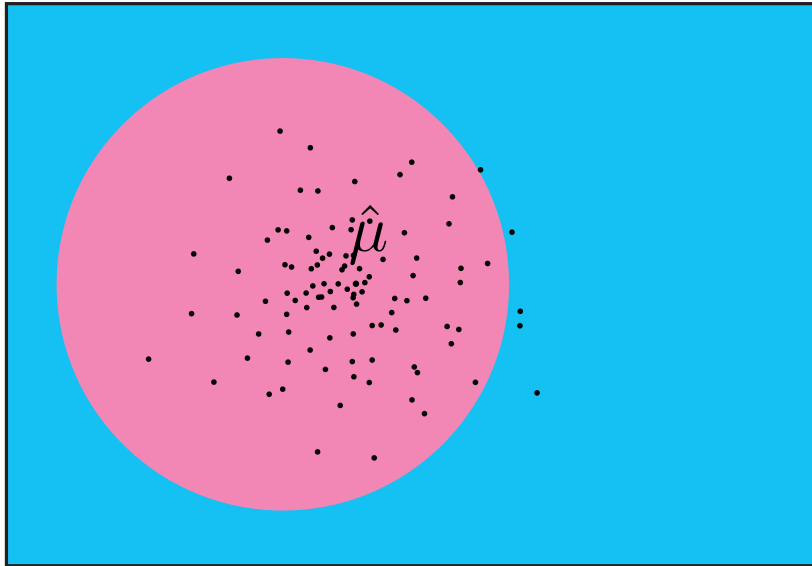


In case 3 - the bootstrap proportion is a good estimate of the P-value
In case 4 - the bootstrap proportion overestimates the P-value

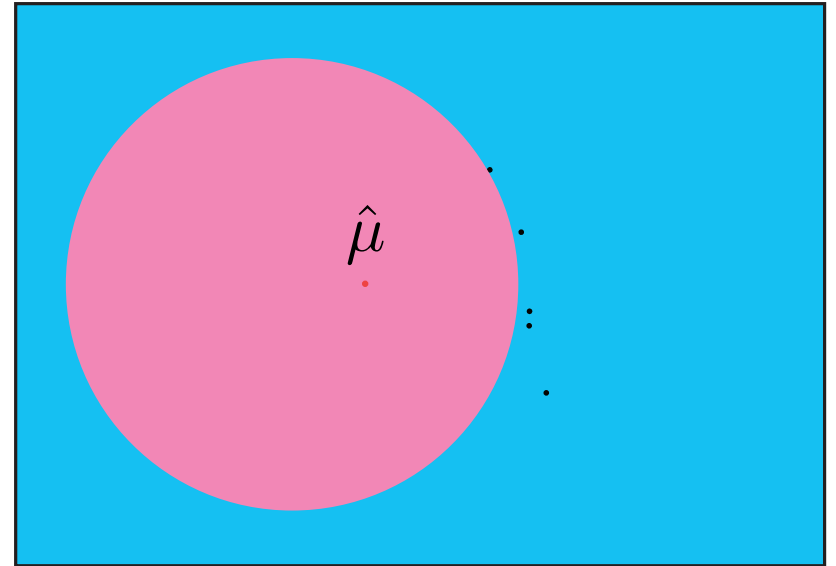
Efron et al. (1996) pointed out these issues of curvature of the boundaries between tree hypotheses.

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.

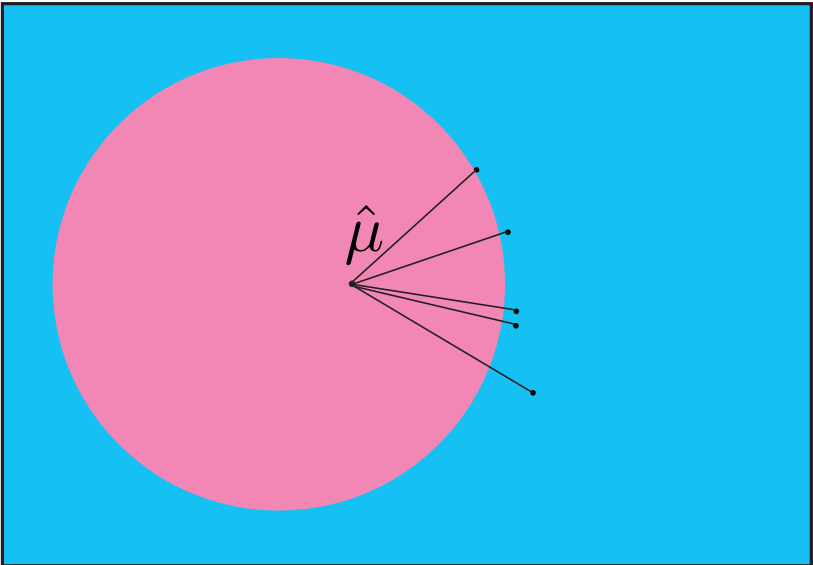
Initial bootstrap



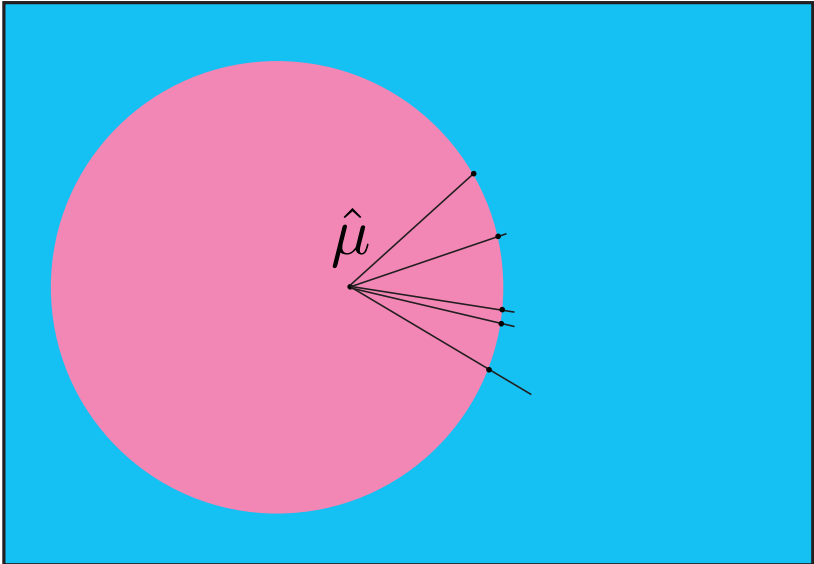
Find replicates that return a tree without the clade



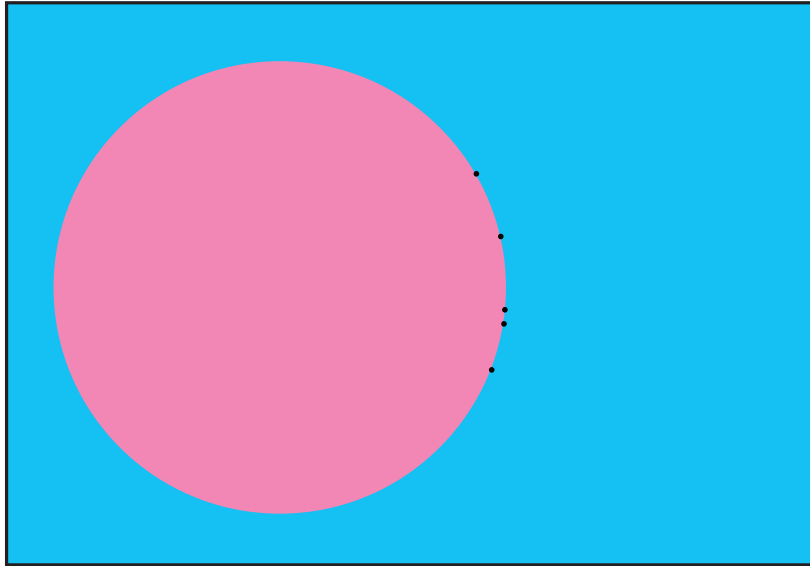
Find replicates that return a tree without the clade



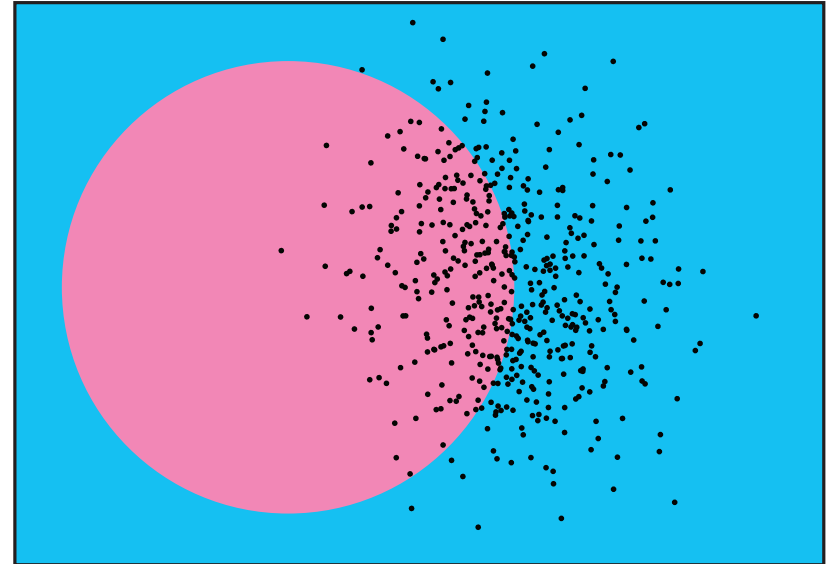
Find boundary points between regions



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:

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

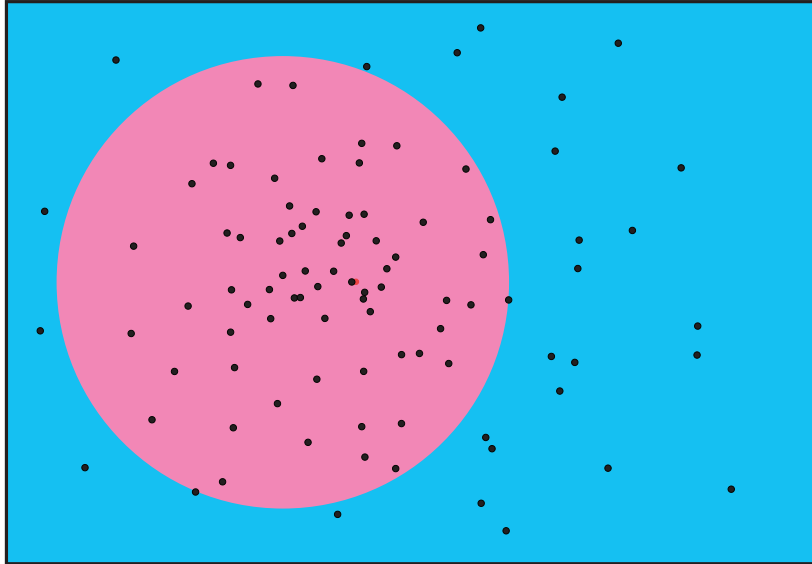
Note the incorrect sign with respect to the curvature term causes BP (and recall how on the curved boundary examples, the curvature caused the P-value to change in one direction and the BP to go in the other).

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

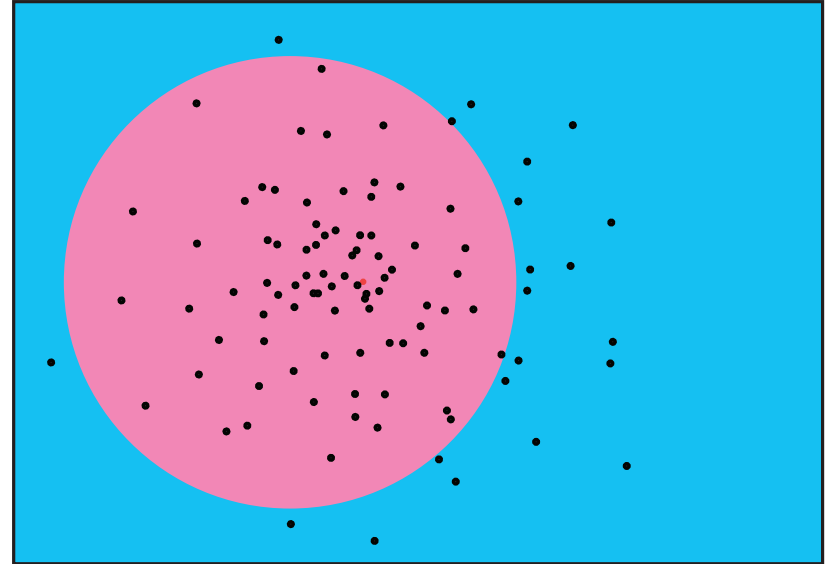
- N is the number of characters in the real data set
- N' is the number of characters in each bootstrap data set
- $r = \frac{N'}{N}$
- If you do a bootstrap in which $r \neq 1$, Shimodaira determined the expected effect on the bootstrap proportion as a function of d and c :

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

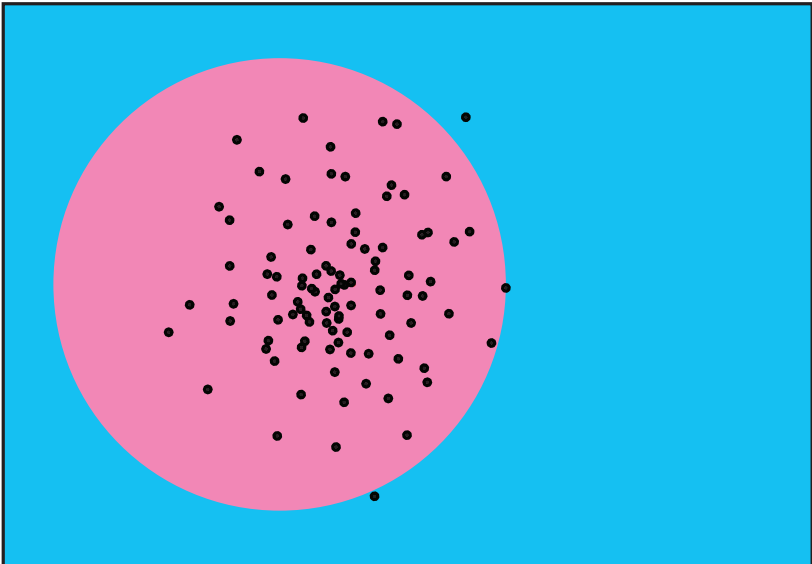
$r = 0.5$



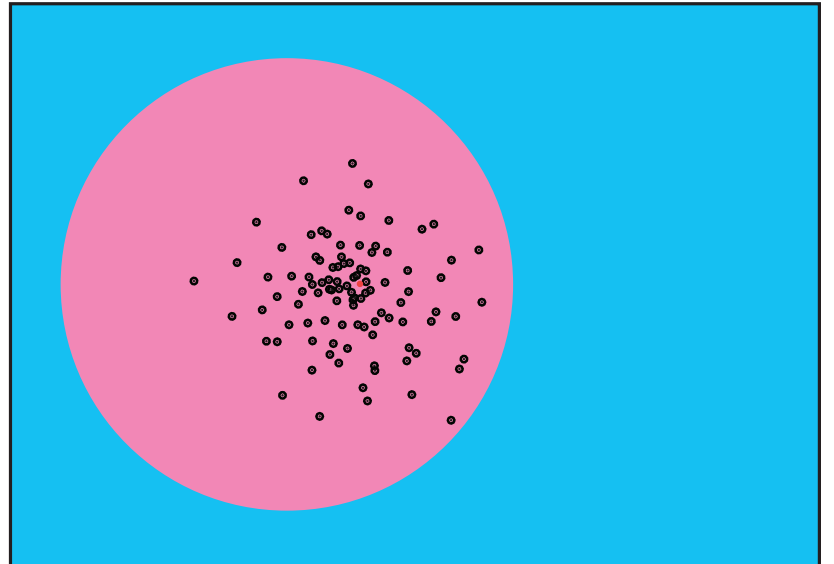
$r = 0.8$



$r = 1.0$



$r = 1.2$



AU Test

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