Some of these slides have been borrowed from Dr. Paul Lewis, Dr. Joe Felsenstein. Thanks!

Paul has many great tools for teaching phylogenetics at his web site:

http://hydrodictyon.eeb.uconn.edu/people/plewis

Simple model selection criteria

Simple model M_1 has k_1 parameters called θ_1, and the complex model M_2 has k_2 parameters called θ_1 ($|\theta_1| = k_1 < k_2 = |\theta_2|$). Data set X has n observations (columns or "characters"):

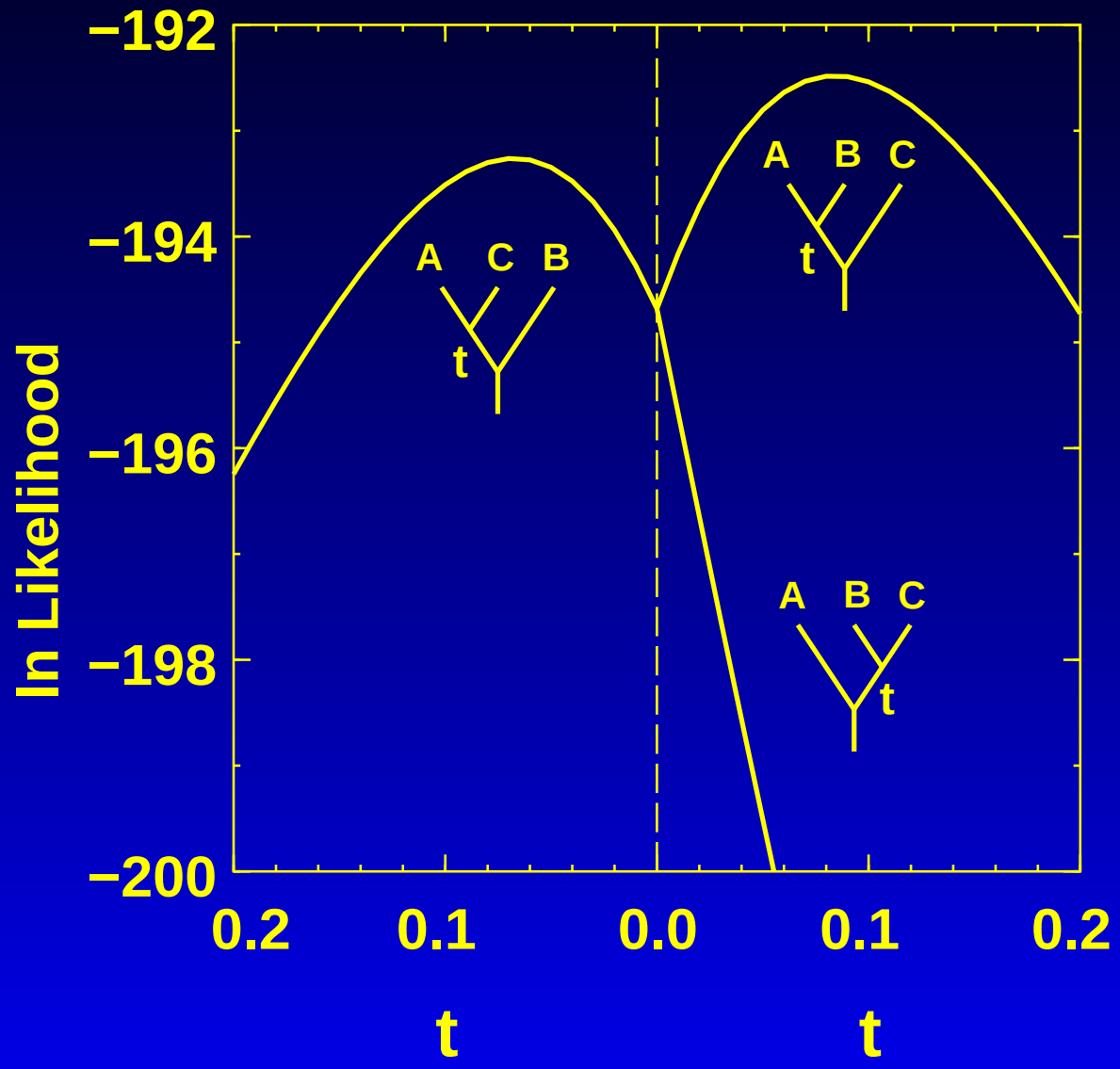
- LRT $= 2\left[\ln \Pr(X|M_2, \widehat{\theta_2}) - \ln \Pr(X|M_1, \widehat{\theta_1})\right]$ and LRT $\sim \chi^2_{k_2-k_1}$ if M_1 is the true model and M_2 is a generalization of M_1 (also stated as M_1 is nested within M_2).

- ΔAIC $= 2\left[\ln \Pr(X|M_1, \widehat{\theta_1}) - \ln \Pr(X|M_2, \widehat{\theta_2})\right] + 2(k_2 - k_1)$

- ΔBIC $= 2\left[\ln \Pr(X|M_1, \widehat{\theta_1}) - \ln \Pr(X|M_2, \widehat{\theta_2})\right] + 2(k_2 - k_1)\ln n$

AIC and BIC prefer the ML tree, but can we use the LRT with the trees as our models?

Can we test trees using the LRT?



Do phylogenetic methods work?

If we collect a data set generated on some evolutionary tree, and then we apply a tree-inference method to the data, then we will get back a tree.

Is the tree that we infer the true tree?

Reasons phylogenetic methods might fail

1. Our data might be mis-scored (garbage-in-garbage-out)
2. Our inference method might not be sophisticated enough. *Systematic error*.
3. We might not have enough data. Sometimes referred to as *random error*.

Are our methods prone to systematic error?

Will we get the wrong tree even if we have enough data?

If so, then we should worry even when we have high statistical support?

Statistical consistency

A popular oversimplification: getting the right answer with infinite amounts of data.

A better definition for trees:

> $\hat{T}(X)$ is the tree estimated by some method, for dataset X
>
> There exists a threshold # of characters, D, such that
>
> $\Pr(\hat{T}(X) = T) > \lambda$
>
> if $|X| > D$, for all $0 \leq \lambda < 1$ and $T \in \mathcal{T}$

This is too general, though...

Statistical consistency (continued)

M is a model of character evolution with parameters, θ.

There exists a threshold $\#$ of characters, D, such that

$\Pr(\hat{T}(X) = T|M, \theta) > \lambda$

if $|X| > D$, for all $0 \leq \lambda < 1, T \in \mathcal{T}$, and $\theta \in \Theta$

Demonstrating consistency

Note that D can be huge, so we can even consider $|X| \to \infty$:

> The proportion of data matrix composed of characters of type $x_i \to \Pr(x_i|M, \theta)$

JC, two taxon data matrix with branch ν:

$$\Pr(\mathbf{X}|\nu) = \begin{pmatrix} \frac{1+3e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} \\ \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1+3e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} \\ \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1+3e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} \\ \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1-e^{\frac{-4\nu}{3}}}{4} & \frac{1+3e^{\frac{-4\nu}{3}}}{4} \end{pmatrix}$$

$$Pr(\text{diff}) = \frac{3 - 3e^{\frac{-4\nu}{3}}}{4}$$

As $n \to \infty$

$$Pr(\text{diff}) = p = \frac{3 - 3e^{\frac{-4\nu}{3}}}{4}$$

recall that under the JC distance correction:

$$d = -\frac{3}{4} \ln \left(1 - \frac{4p}{3} \right)$$

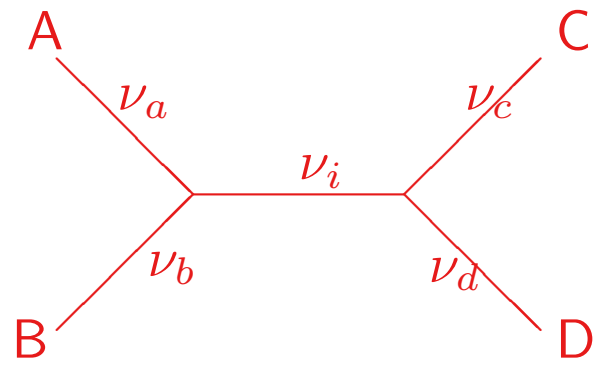
so as $n \to \infty \ldots$

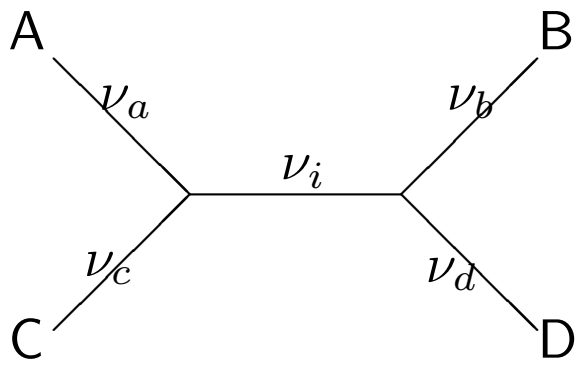
$$
\begin{aligned}
d &= -\frac{3}{4}\ln\left(1-\frac{4\left[\frac{3-3e^{\frac{-4\nu}{3}}}{4}\right]}{3}\right) && (1)\\
&= -\frac{3}{4}\ln\left(1-\frac{3-3e^{\frac{-4\nu}{3}}}{3}\right) && (2)\\
&= -\frac{3}{4}\ln\left(1-1+e^{\frac{-4\nu}{3}}\right) && (3)\\
&= -\frac{3}{4}\frac{-4\nu}{3} && (4)\\
&= \nu && (5)
\end{aligned}
$$

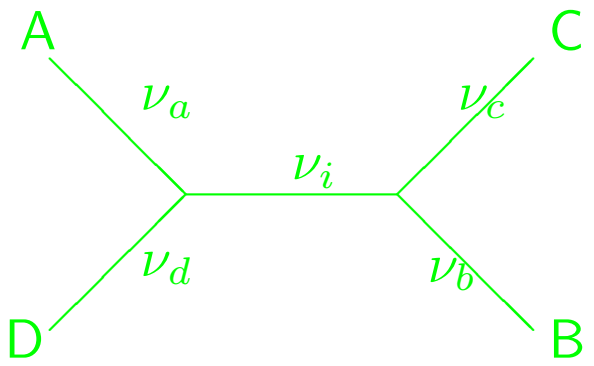
Thus, as $n \to \infty$, $d_{ij} \to \nu_{ij}$. In other lectures we have referred to the "branch length" between tips, ν_{ij} as the path length p_{ij}.

This means that if we correct distances using the correct model, the corrected distances will converge to the true path lengths.

Buneman's logic proves that if all pairwise distances are correct then we can infer the tree (only one tree will be capable of having a set of edge lengths that are additive.).



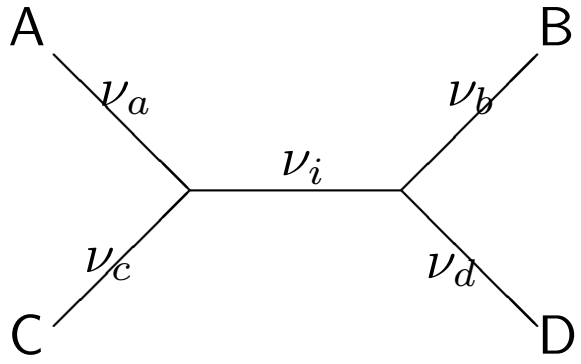




$d_{AB} + d_{CD}$	$\nu_a + \nu_b + \nu_c + \nu_d$	$\nu_a+\nu_b+\nu_c+\nu_d+2\nu_i$	$\nu_a + \nu_b + \nu_c + \nu_d + 2\nu_i$
$d_{AC} + d_{BD}$	$\nu_a+\nu_b+\nu_c+\nu_d+2\nu_i$	$\nu_a + \nu_b + \nu_c + \nu_d$	$\nu_a + \nu_b + \nu_c + \nu_d + 2\nu_i$
$d_{AD} + d_{BC}$	$\nu_a+\nu_b+\nu_c+\nu_d+2\nu_i$	$\nu_a+\nu_b+\nu_c+\nu_d+2\nu_i$	$\nu_a + \nu_b + \nu_c + \nu_d$

The four point condition of Buneman (1971).

This assumes additivity of distances.



$d_{AB} + d_{CD}$	$\nu_a + \nu_b + \nu_c + \nu_d + 2\nu_i + \epsilon_{AB} + \epsilon_{CD}$
$d_{AC} + d_{BD}$	$\nu_a + \nu_b + \nu_c + \nu_d + \epsilon_{AC} + \epsilon_{BD}$
$d_{AD} + d_{BC}$	$\nu_a + \nu_b + \nu_c + \nu_d + 2\nu_i + \epsilon_{AD} + \epsilon_{BC}$

If $|\epsilon_{ij}| < \frac{\nu_i}{2}$ then $d_{AC} + d_{BD}$ will still be the smallest sum – So Buneman's method will get the tree correct.

Worst case: $\epsilon_{AC} = \epsilon_{BD} = \frac{\nu_i}{2}$ and $\epsilon_{AB} = \epsilon_{CD} = -\frac{\nu_i}{2}$ then

$$d_{AC} + d_{BD} = \nu_a + \nu_b + \nu_c + \nu_d + \nu_i = d_{AB} + d_{CD}$$

Reasons phylogenetic methods might fail

1. Our data might be mis-scored (garbage-in-garbage-out)
2. Our inference method might not be sophisticated enough. *Systematic error*.
3. We might not have enough data. Sometimes referred to as *random error*.

Is our data set large enough to avoid random errors?

Assessing the support for our estimates:

1. Topology tests (KH test, PTP...)

2. Bootstrapping,

3. Model-based simulation tests

Frequentist phylogenetic hypothesis testing?

If we have a tree, T_0, in mind *a priori*, then how can we answer the question:

"Based on this data, should we reject the tree?"

Clearly if $\hat{T} \neq T_0$, then there is a possibility that we should reject T_0.

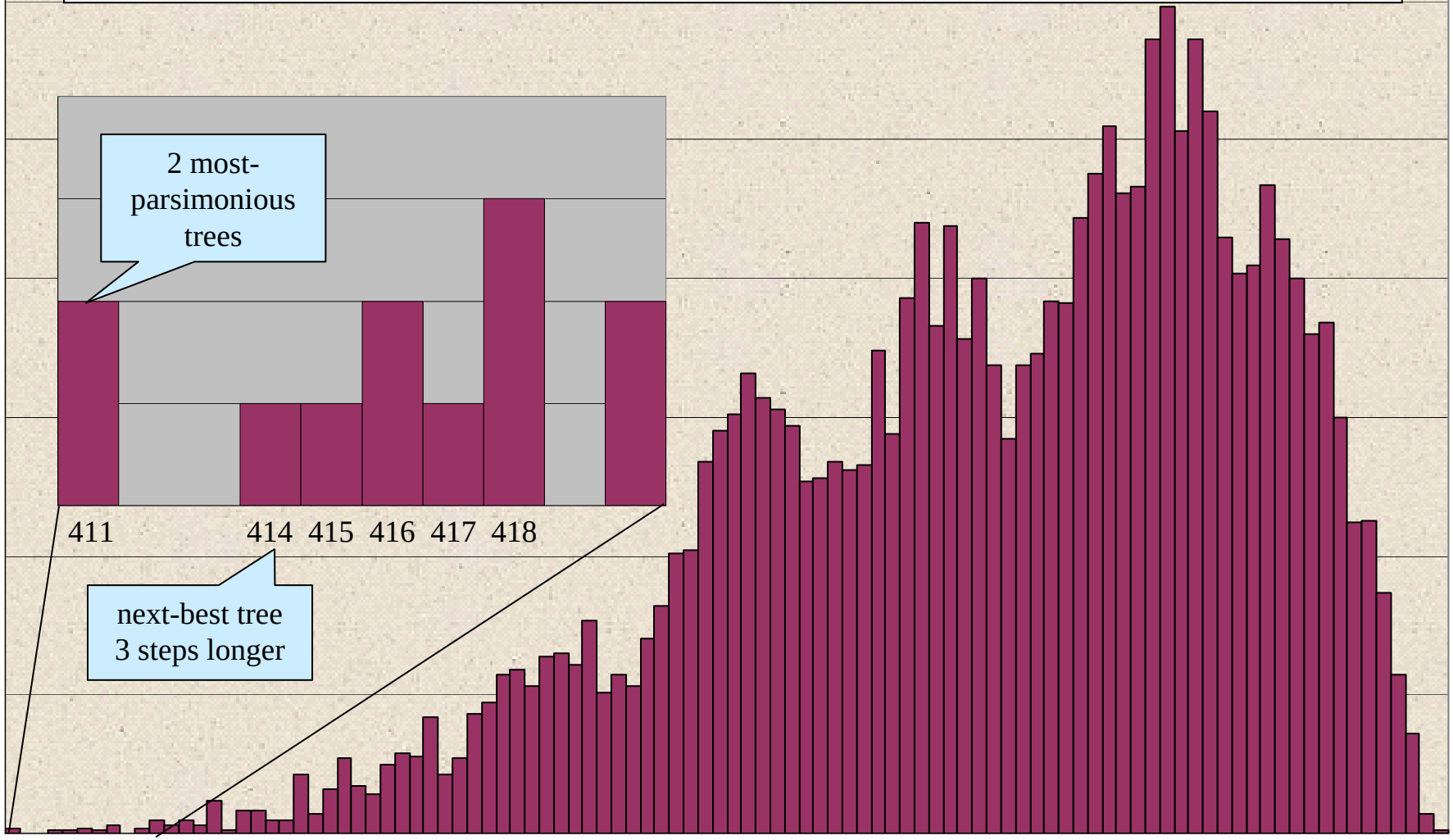
But how do we calculate a p-value?

Frequentist phylogenetic hypothesis testing? (cont)

What is our test-statistic?

How do we find the null-distribution for the test statistic?

Exhaustive search: algae.nex



Frequentist phylogenetic hypothesis testing? (cont)

What is our test-statistic? Difference in support between the null tree and the preferred tree.

How do we find the null-distribution for the test statistic?

1. Search for the preferred tree, $\hat{T}$
2. Search for the best tree that is consistent with the null hypothesis, T_0
3. Let z be the difference in score (parsimony, ME, SSE, lnL) between these trees.

How big does z have to be in order for us to reject the null hypothesis?

Frequentist phylogenetic hypothesis testing? (cont)

What is our test-statistic? Difference in support between the null tree and the preferred tree.

How do we find the null-distribution for the test statistic? Simulation

1. Simulate a large number of data sets on T_0. On each dataset, i:
   (a) Search for the preferred tree, $\hat{T}^{(i)}$
   (b) Search for the best tree that is consistent with the null hypothesis, $T_0^{(i)}$
   (c) Let $z_0^{(i)}$ be the difference in score between these trees for data set i
2. See if the observed test statistic z is in the $\alpha\%$ tail of the distribution $\mathbf{z}_0$

Testing 2 trees

Test statistic: The difference in score, δ, between two trees chosen *a priori*.

Null hypothesis: Both trees are equally good explanations of the truth.

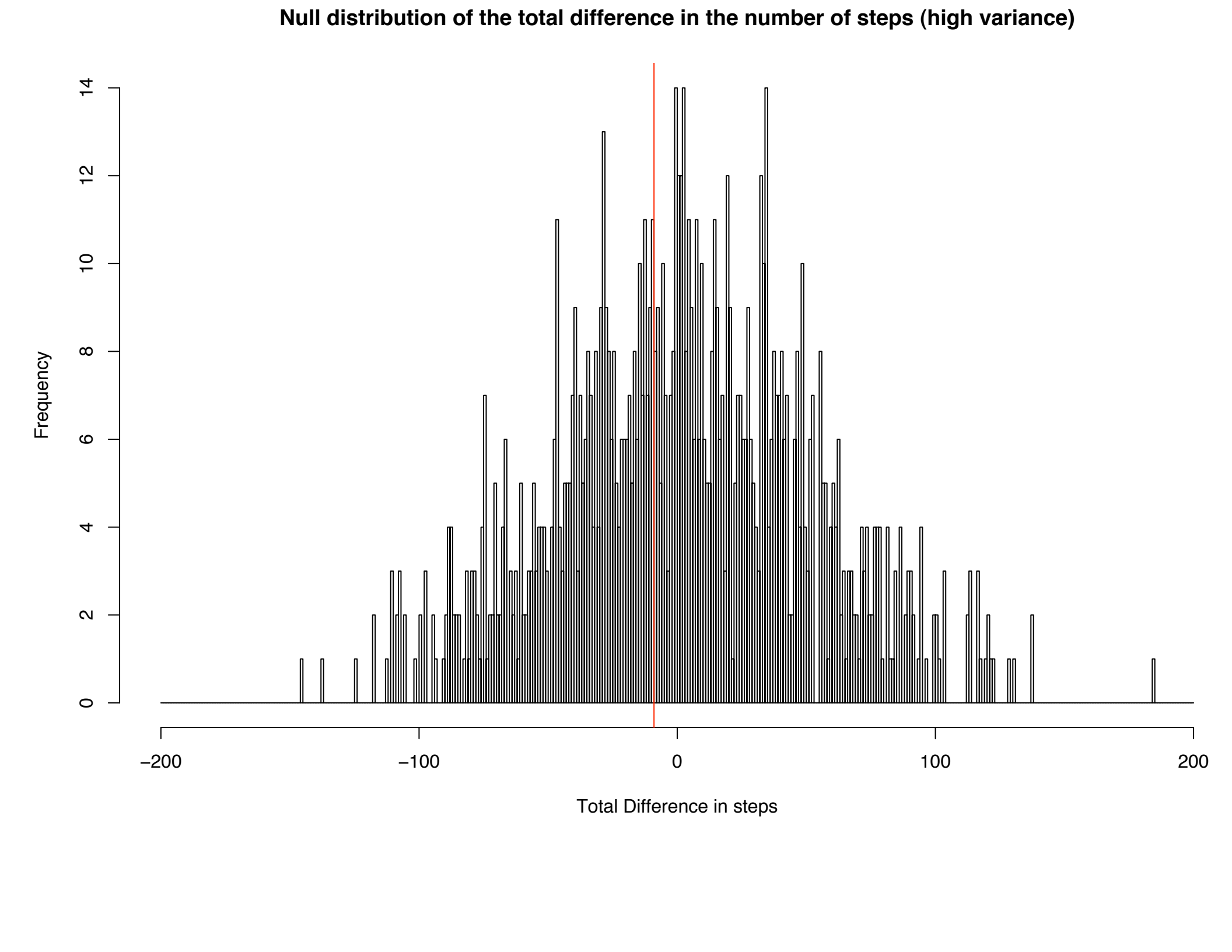
$$E(\delta) = 0$$

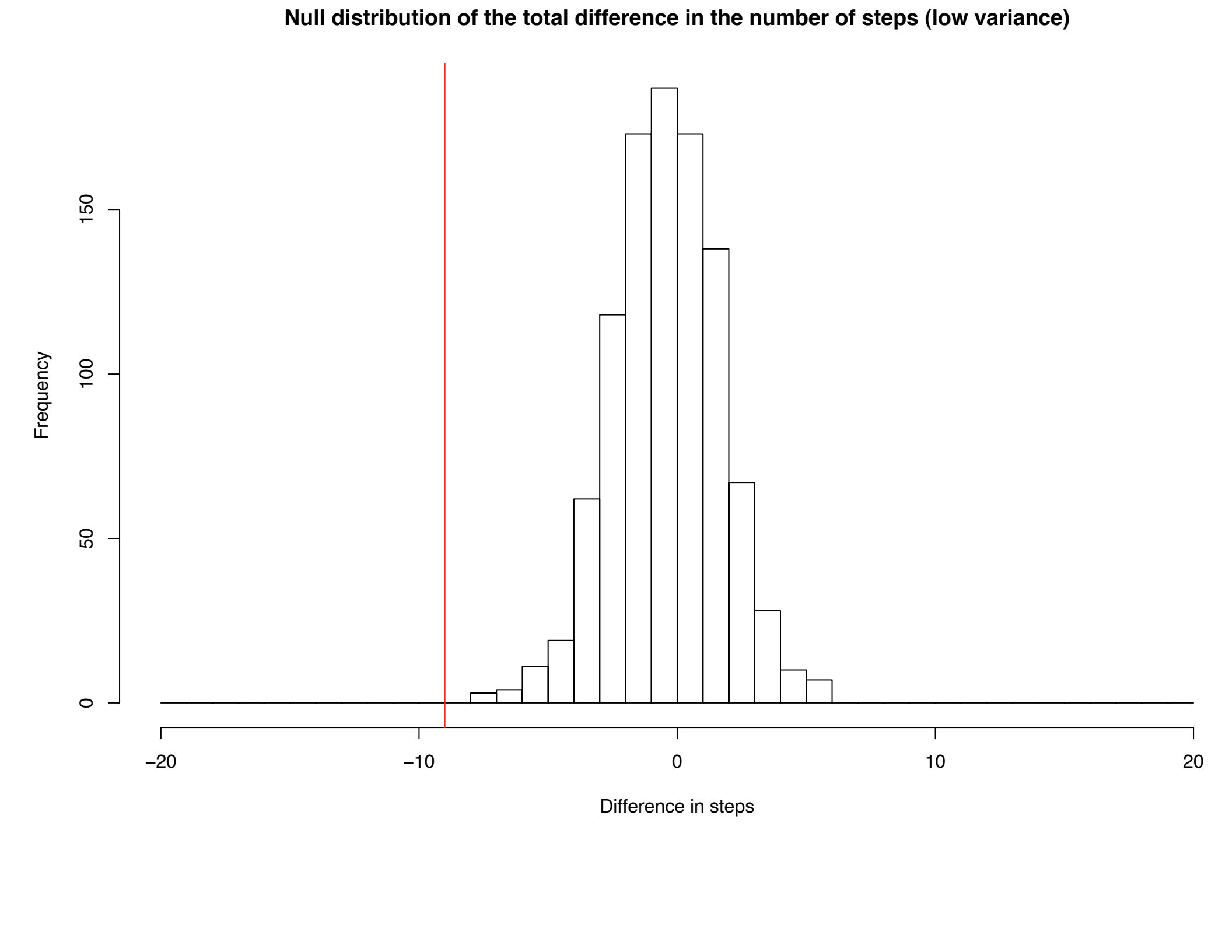
Is δ so large that we reject the null?
What if $\delta = -9$, should we reject the null?

Tests that treat the number of variable length characters as fixed

1. Winning sites test (test the null that there is an equal probability of a site favoring tree 1 over tree 2)
2. Templeton's test a version of Wilcoxon's rank sum test

Robust, but not very powerful


Null distribution of the total difference in the number of steps (high variance)
Frequency
0
2
4
6
8
10
12
14
−200
−100
0
100
200
Total Difference in steps

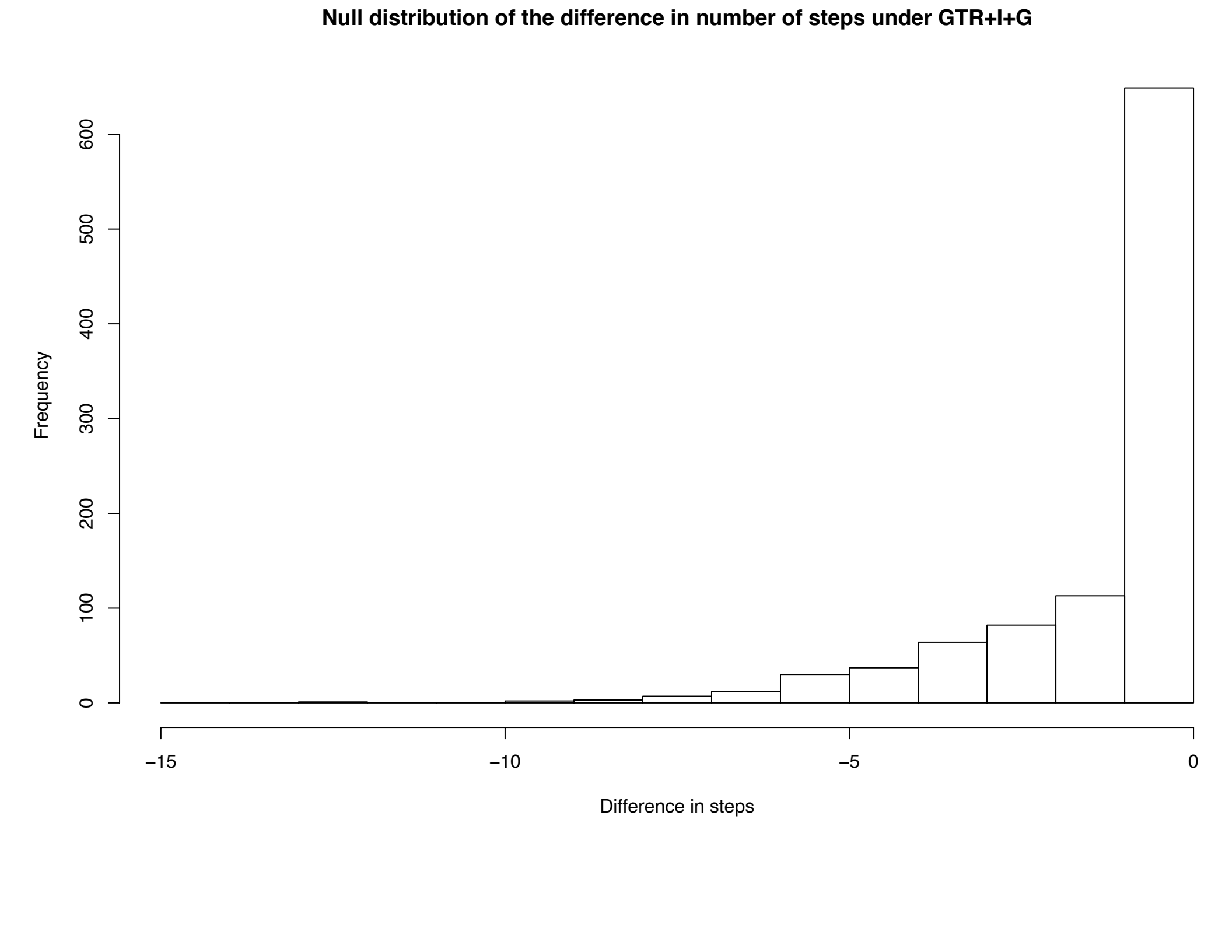

Null distribution of the total difference in the number of steps (low variance)
Frequency
0
50
100
150
−20
−10
0
10
20
Difference in steps

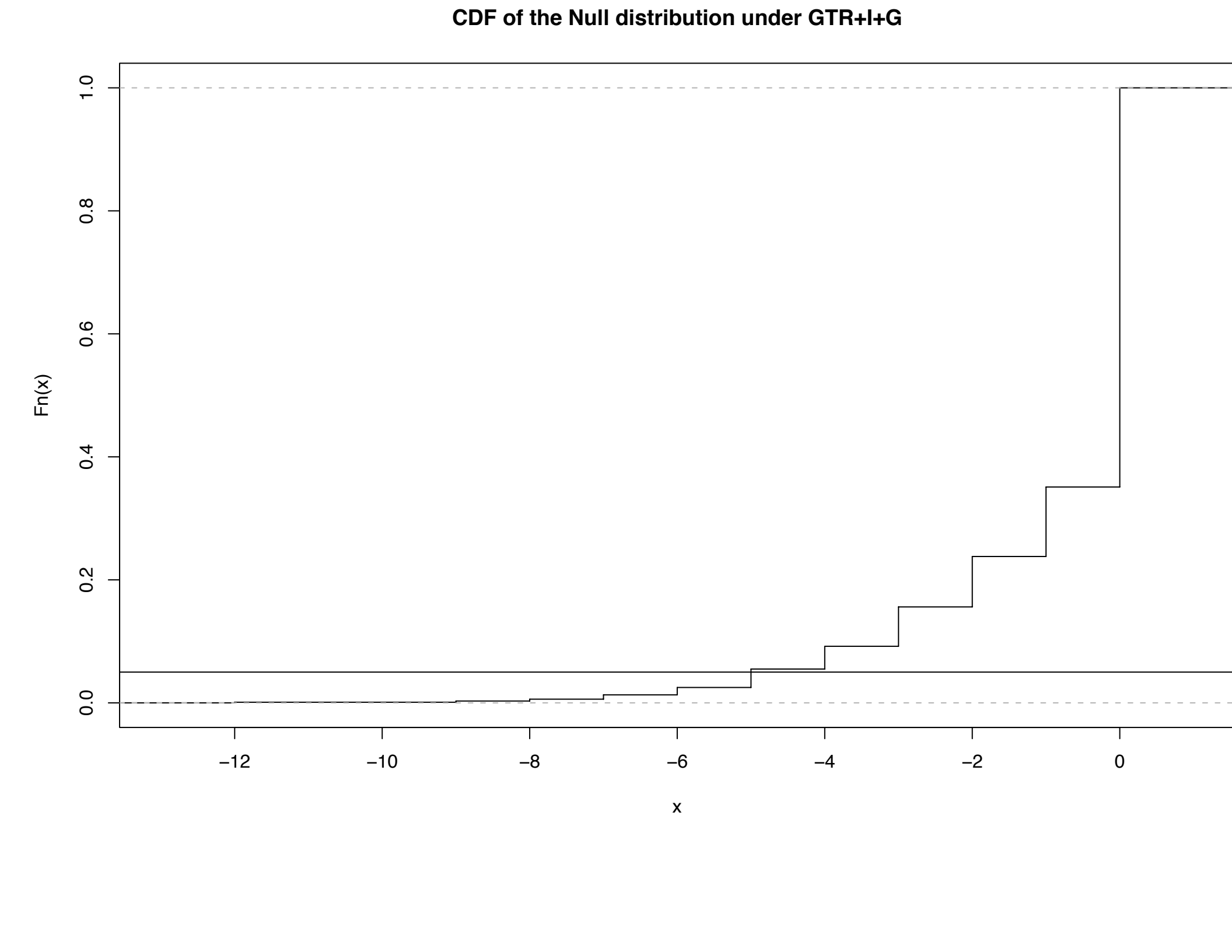
I could generate the last 4 slides because I made up (and hence knew) a variance.

How can we assess the strength of support if we do not know the variance of the generating process?

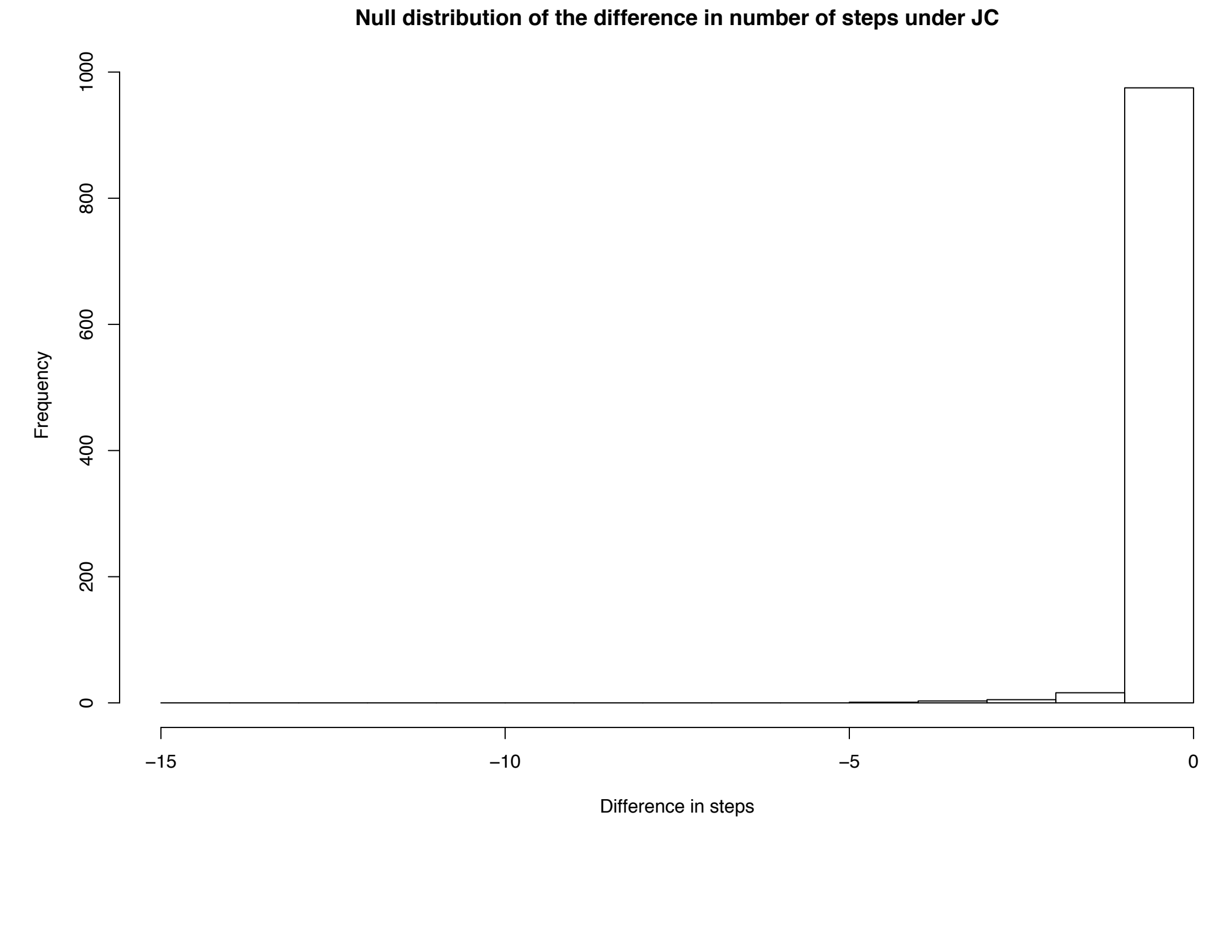
Parametric bootstrapping

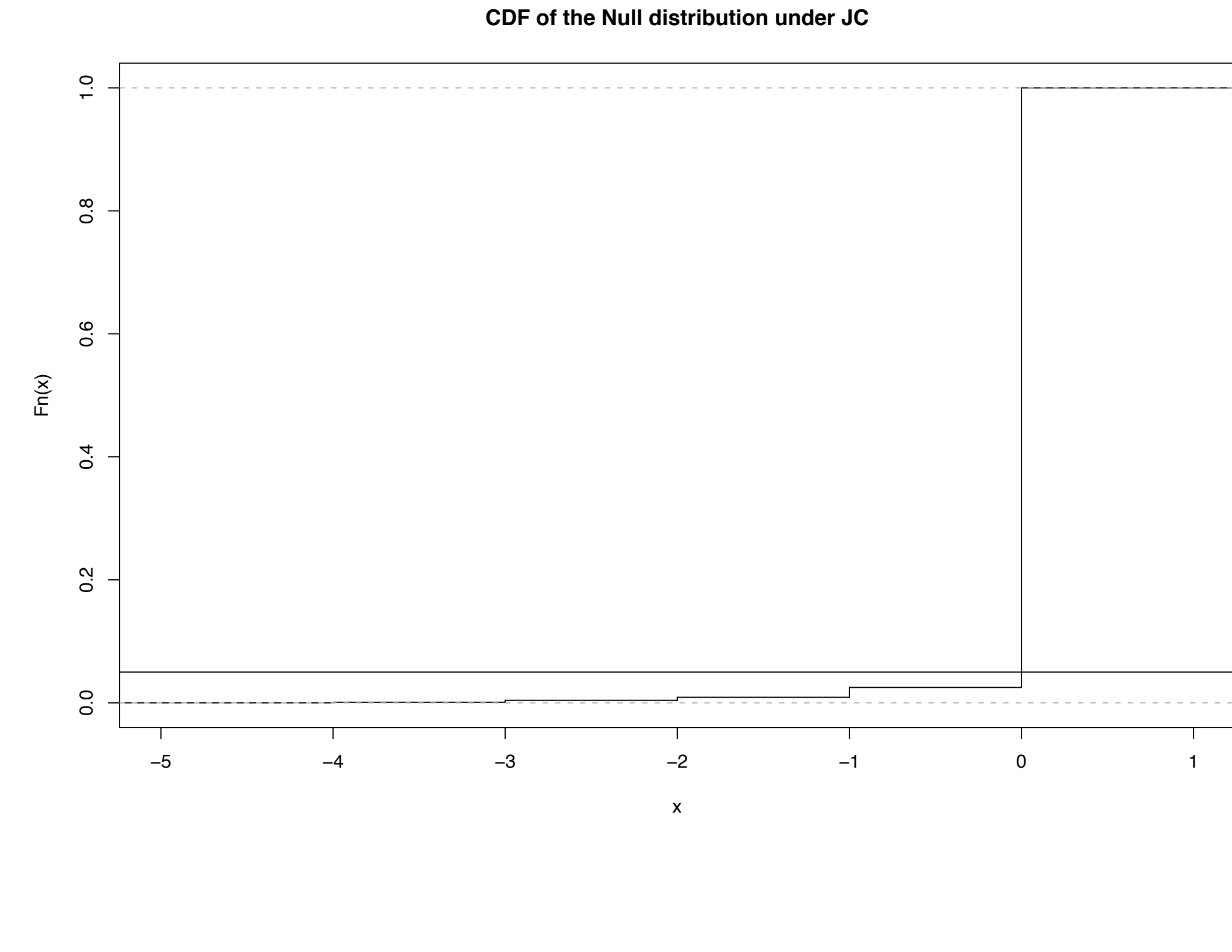
1. Simulate a large number of data sets on T_0. On each dataset, i:
   (a) Search for the preferred tree, $\hat{T}^{(i)}$
   (b) Search for the best tree that is consistent with the null hypothesis, $T_0^{(i)}$
   (c) Let $z_0^{(i)}$ be the difference in score between these trees for data set i
2. See if the observed test statistic z is in the $\alpha\%$ tail of the distribution $\mathbf{z}_0$


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


CDF of the Null distribution under GTR+I+G
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−12
−10
−8
−6
−4
−2
0
x

Parametric bootstrapping is powerful, but can also be sensitive to the model chosen to generate the data.

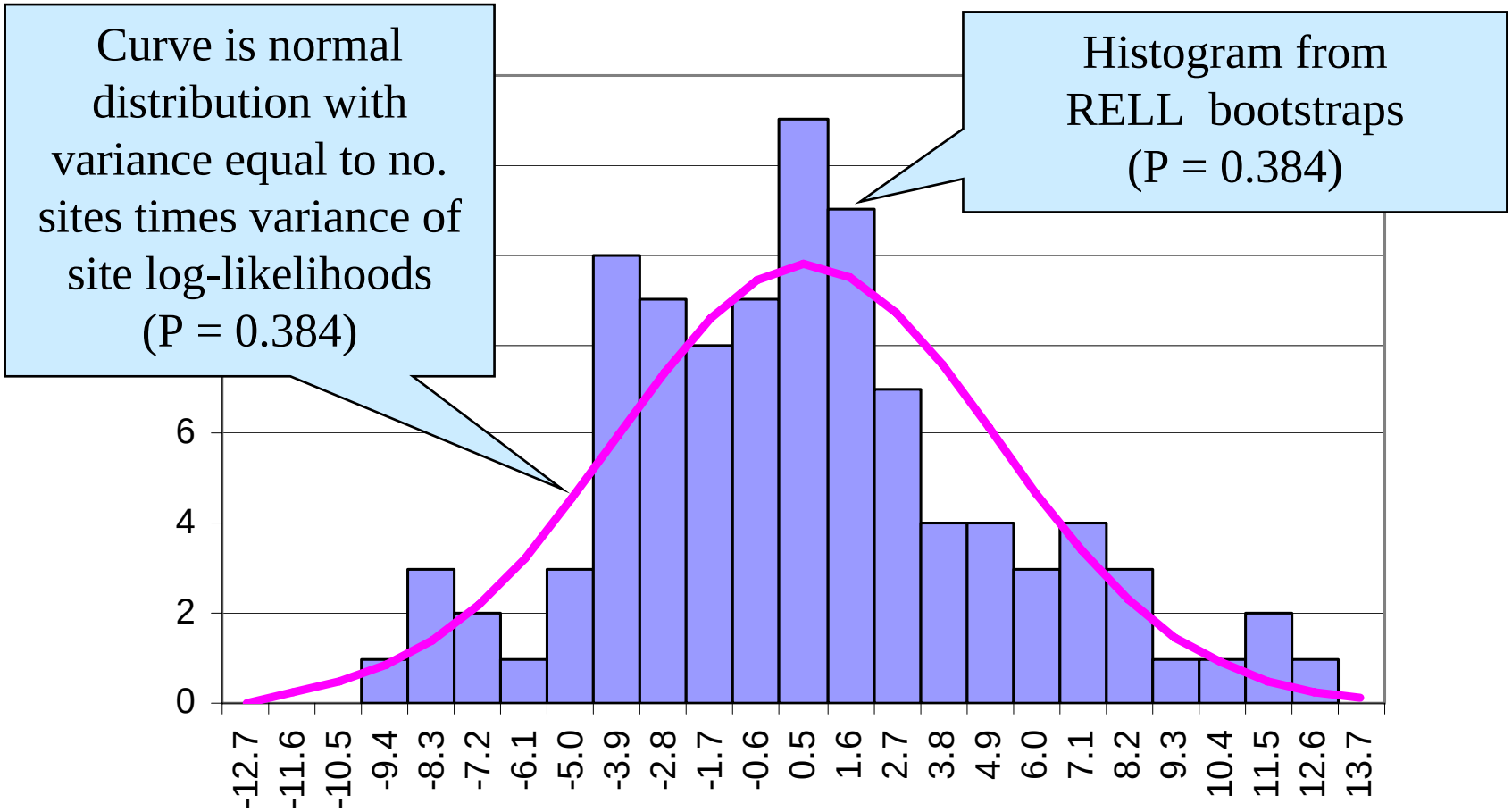

Null distribution of the difference in number of steps under JC
Frequency
0
200
400
600
800
1000
−15
−10
−5
0
Difference in steps


CDF of the Null distribution under JC
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−5
−4
−3
−2
−1
0
1
x

Kishino Hasegawa Test

1. Several variants:
   (a) parametric - using a normal approximation
   (b) non-parametric - using a bootstrapping
2. Appropriate for testing the null hypothesis that two trees explain the data equally well.
3. **Both** trees for the test must be specified from prior knowledge.
4. Use the site-to-site variation in δ to estimate the variance of a Normal distribution
5. In the null the mean is 0

KH (Normal Approx.) Test

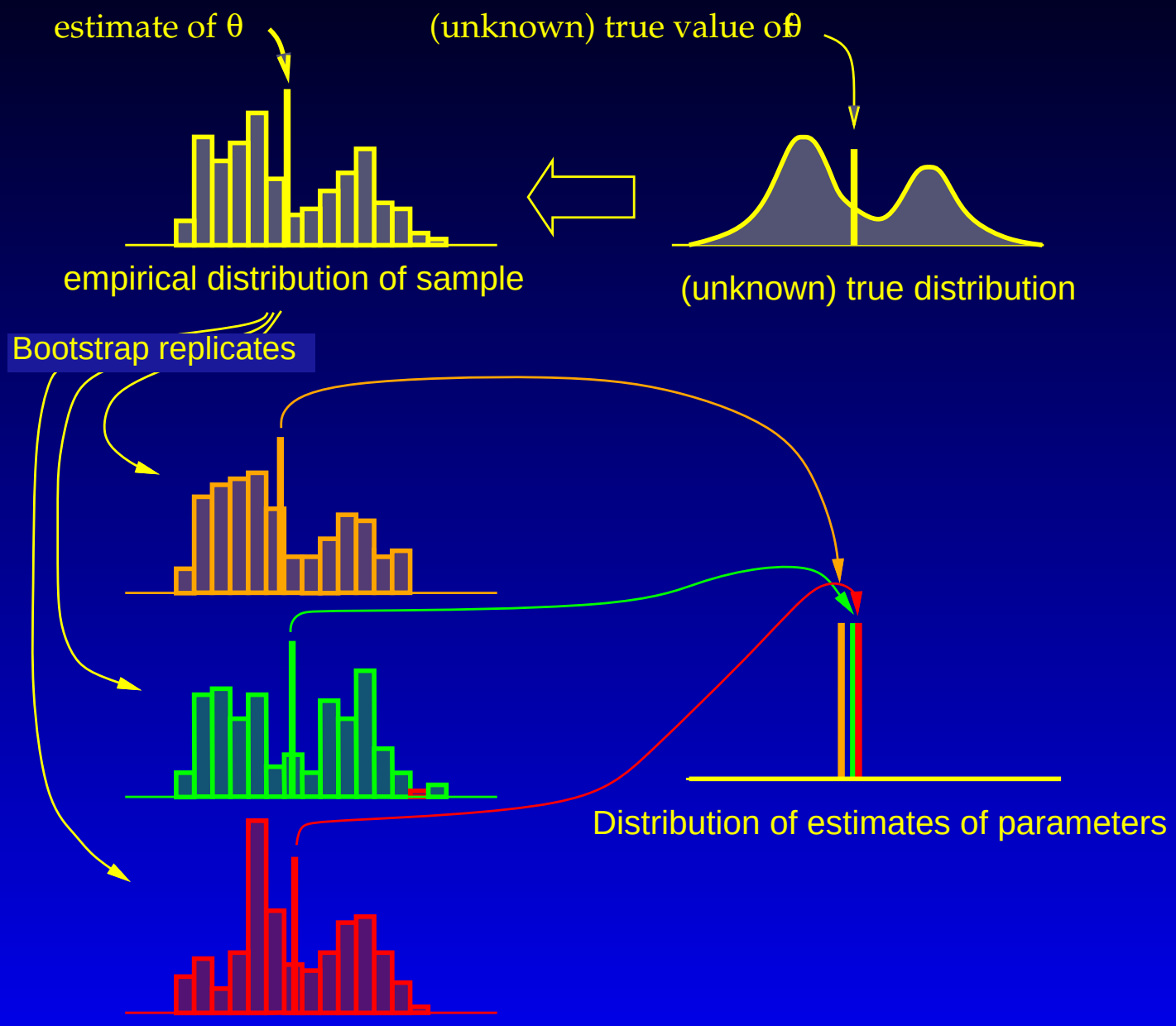


Kishino Hasegawa Test (bootstrapping)

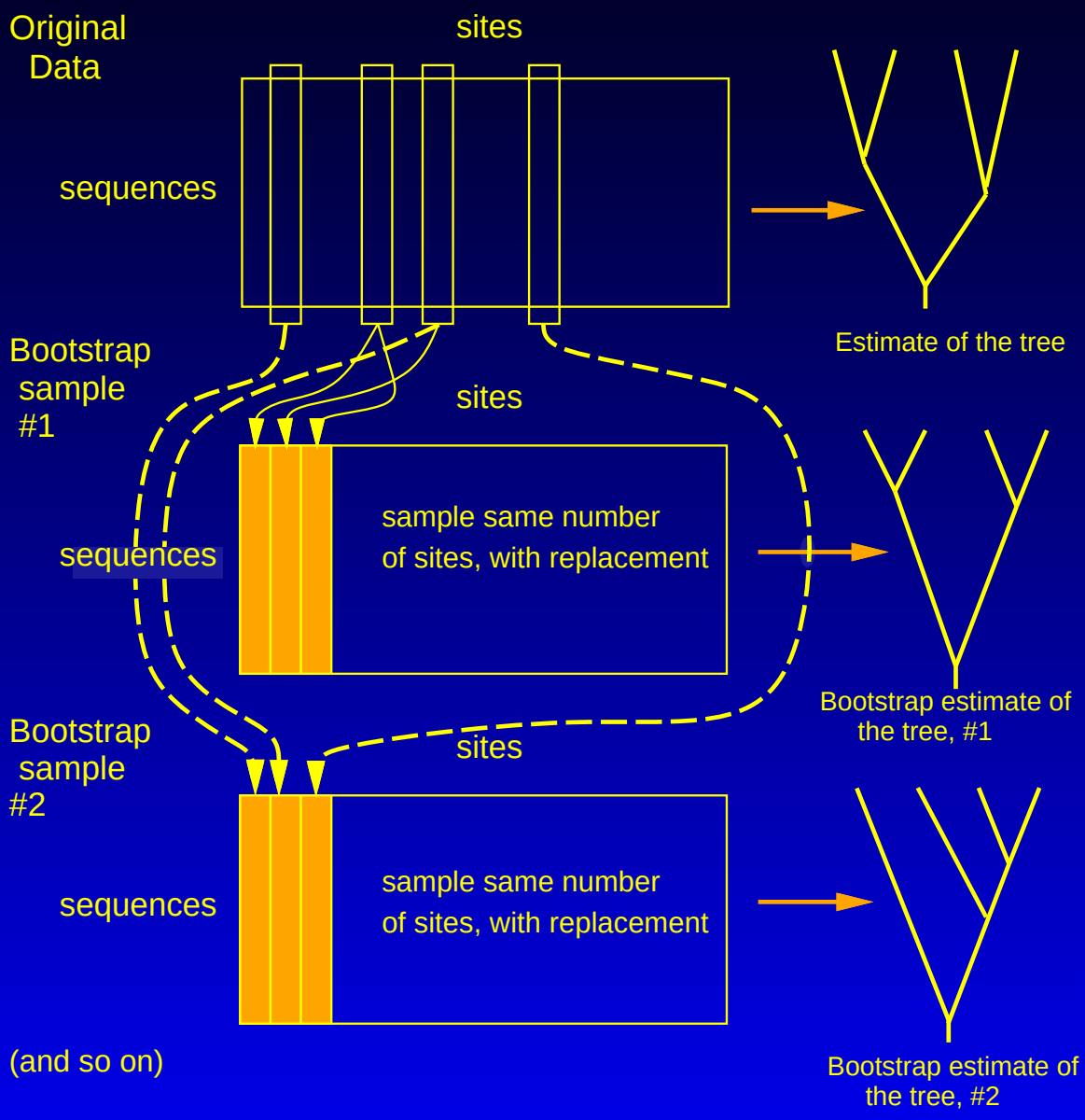
Rather than assume a Normal distribution and estimate a variance, it is common to bootstrap the data to generate a null distribution of the total difference in score between 2 trees.

Bootstrapping is resampling your data to mimic the variability in the process that generated the data.

The bootstrap



The bootstrap for phylogenies



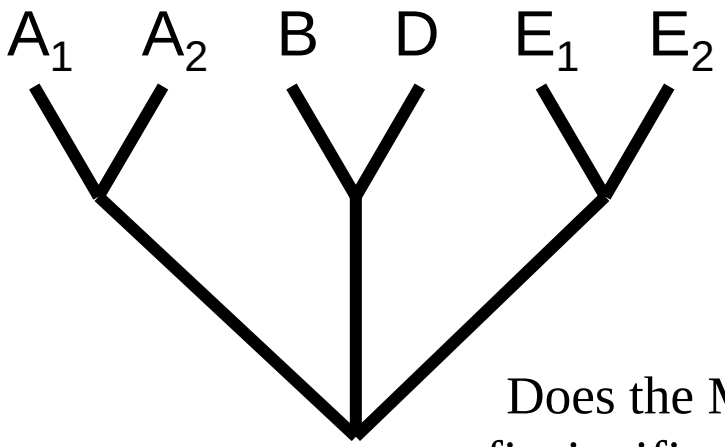
Kishino Hasegawa Test (bootstrapping - continued)

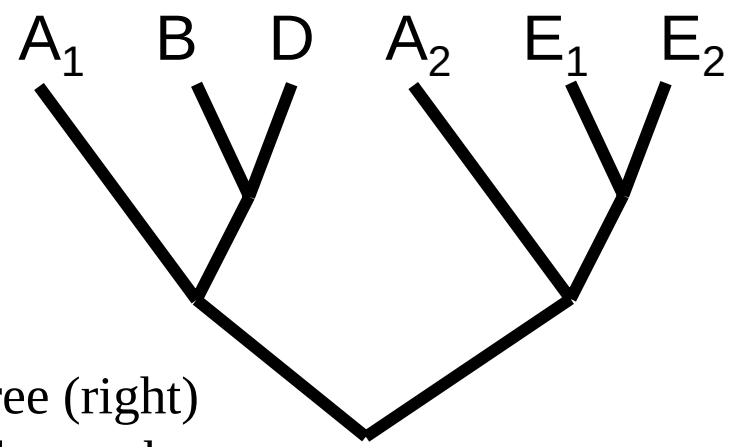
Bootstrapping can give us a sense of the variability, but if the original data favored tree 1 isn't it very likely that the bootstrapped replicates will be more likely to support tree 1 than tree 2?

This does not sound like we are following our null that both trees are equally good.

Solution: we must center the bootstrapped differences in score.

Example: HIV-1 subtypes





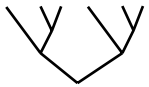
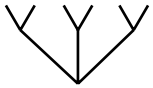
$\ln L_1 = -5073.75$

Does the ML tree (right) fit significantly better than the accepted phylogeny (left)?

$\ln L_{ML} = -5069.85$

2,000 nucleotide sites from *gag* and *pol* genes. Substitution model: GTR+Γ

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

KH test			δ	δ (centered)
Original dataset	-5069.85	- (-5073.75)	= 3.90	---
Bootstrap 1	-4951.07	- (-4958.72)	= 7.65 - 4.43 =	3.23
Bootstrap 2	-5149.91	- (-5158.69)	= 8.78 - 4.43 =	4.35
Bootstrap 3	-5100.88	- (-5104.89)	= 4.01 - 4.43 =	-0.42
● ● ●				
Bootstrap 100	-5051.14	- (-5057.16)	= 6.02 - 4.43 =	1.59

4.43 <- bootstrap mean

KH Test (RELL method)

- Maximize log-likelihood on both topologies, T_1 and T_2
- Calculate $\delta_0 = \ln L_1 - \ln L_2$
- Create n_{reps} (e.g. 100) bootstrap datasets
- Compute δ_i for each bootstrap dataset *i* (this is the RELL part) using the parameter values estimated from the *original data* on T_1 and T_2
- Subtract the mean from each δ_i value so that the mean of the distribution of δ_i values is 0.0
- Compare δ_0 to this distribution. If it falls in the upper or lower 2.5% tail, then reject the null hypothesis that both topologies are equally well supported by the data
- RELL means "Resampling Estimated Log-Likelihood" and refers to using the original parameter estimates rather than reestimating branch lengths and other model parameters for each bootstrap data set separately

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution 29:170-179.

Shimodaira-Hasegawa (SH) Test

- KH test inappropriate when the ML tree is included amongst the trees to test
- SH test corrects for the fact that we know that the ML tree is the best supported
- Like KH, SH test is nonparametric
- One-tailed test because it is based on differences that must all be positive (each difference is $\ln L_{ML} - \ln L_i$ for topology *i*
- Null hypothesis is that all included topologies (may be more than 2) are equally well supported by the data

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16: 1114-1116.

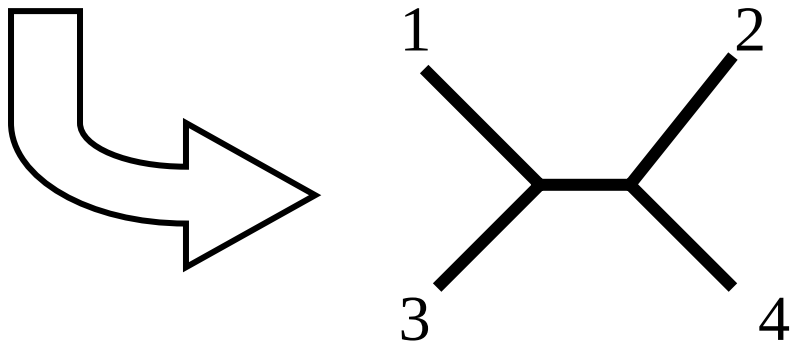
SH Test

1. Calculate the difference in lnL between each tree and the ML tree. Call this δ_i
2. Score each tree in your set on each bootstrap pseudoreplicate data set (using full optimization or RELL)
3. Center the set of scores for each tree - force each tree's the mean lnL to be 0.0 by subtraction.
4. For each bootstrap replicate calculate the difference between the best centered lnL and each tree's lnL - this is the $\delta_{i,j}$ for tree i and bootstrap replicate j.
5. For each tree count the proportion of bootstrap replicates in which δ_{ij} is larger (more extreme) than δ_i. This is the p-value for the tree.
6. Reject trees that have p-values below the significance level for your test.

Bootstrapping: first step

	1	2	3	4	5	6	7	...	*k*
1	T	A	G	T	C	G	T	...	A
2	T	C	A	T	C	G	T	...	G
3	A	T	G	T	C	A	C	...	G
4	A	T	A	T	C	G	C	...	G

From the original data, estimate a tree using, say, parsimony (could use NJ, LS, ML, etc., however)

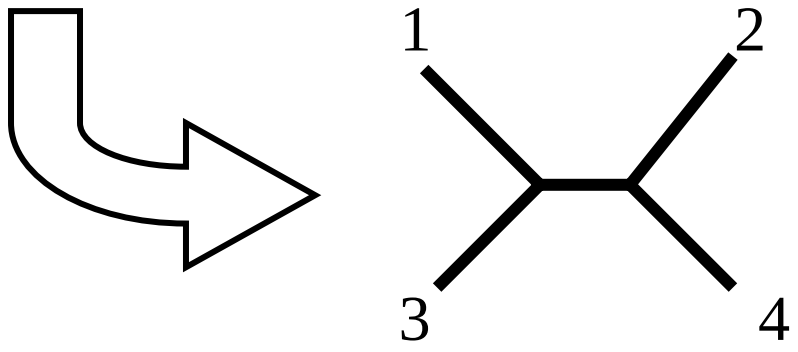


Bootstrapping: first replicate

	1	2	3	4	5	6	7	...	*k*
weights	1	2	0	0	1	3	1	...	2
1	T	A	G	T	C	G	T	...	A
2	T	C	A	T	C	G	T	...	G
3	A	T	G	T	C	A	C	...	G
4	A	T	A	T	C	G	C	...	G

Sum of weights equals *k* (i.e., each bootstrap dataset has same number of sites as the original)

From the bootstrap dataset, estimate the tree using the same method you used for the original dataset

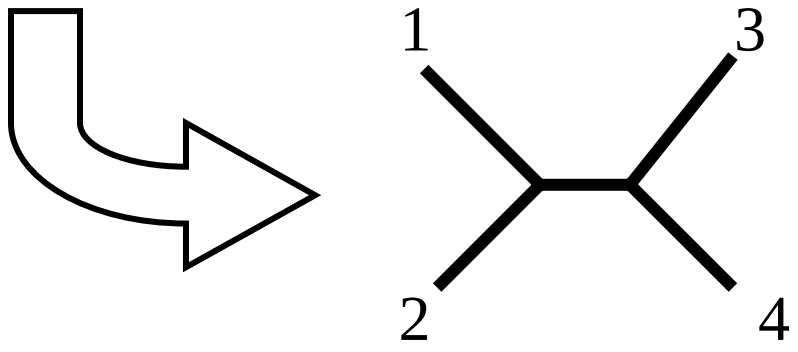


Bootstrapping: second replicate

	1	2	3	4	5	6	7	...	*k*
weights	0	1	1	1	1	3	0	...	0
1	T	A	G	T	C	G	T	...	A
2	T	C	A	T	C	G	T	...	G
3	A	T	G	T	C	A	C	...	G
4	A	T	A	T	C	G	C	...	G

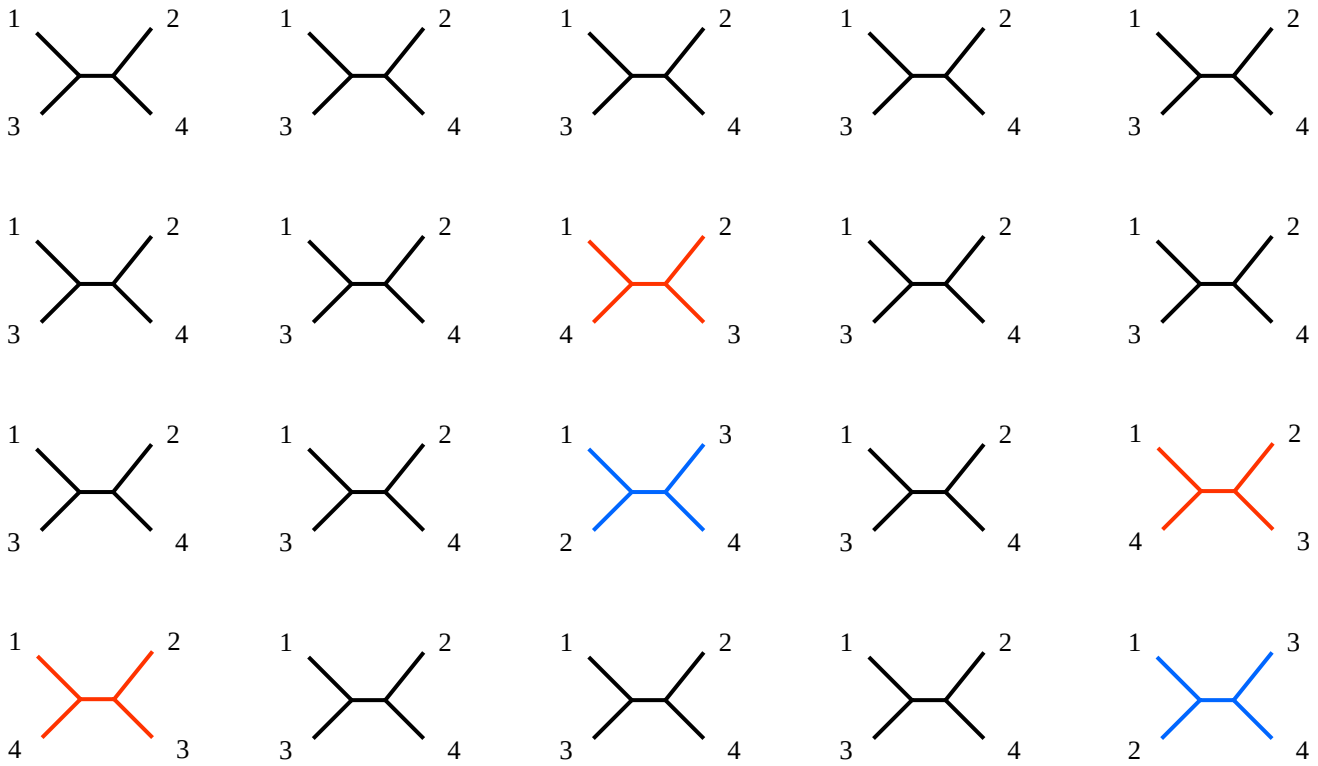
Note that weights are different this time, reflecting the random sampling with replacement used to generate the weights

This time the tree that is estimated is different than the one estimated using the original dataset.



Bootstrapping: 20 replicates

1234	Freq
-*-*	75.0
-**-	15.0
--**	10.0



Note: usually at least 100 replicates are performed, and 500 is better