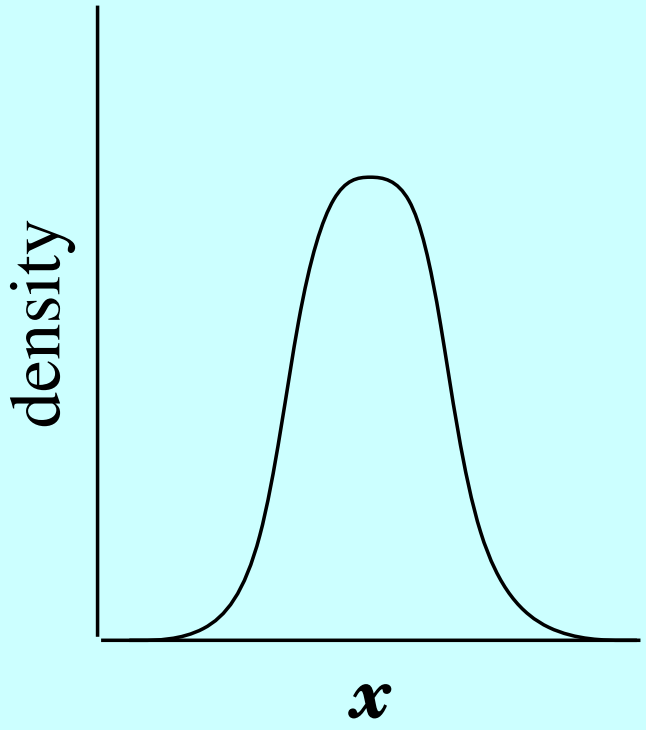
Week 8: Testing trees, Bootstraps, jackknifes, gene frequencies

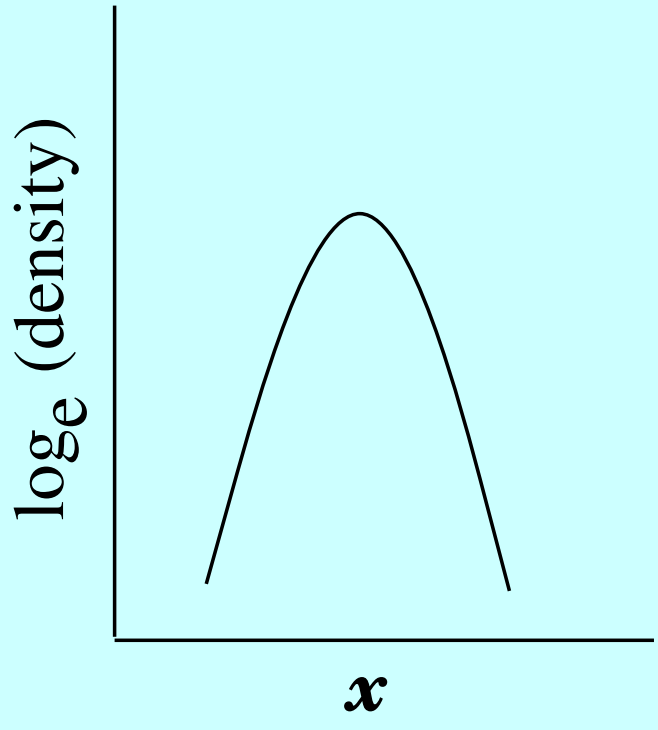
Genome 570

February, 2016

Normal distribution: curvature of log of height



$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

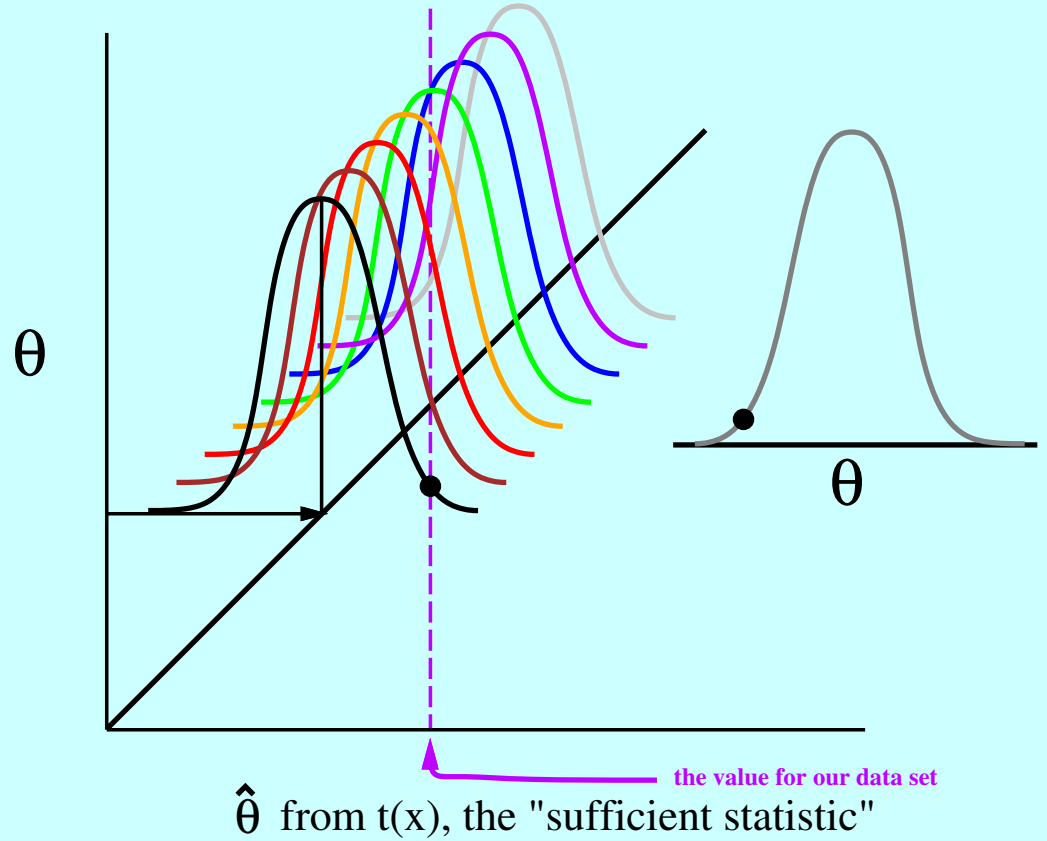


$$(\text{constant stuff}) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

Taking the logarithm of the height of the density curve of a normal distribution whose variance is σ^2, we see that it is a quadratic curve whose curvature is $-1/\sigma^2$

The likelihood curve is nearly a normal distribution

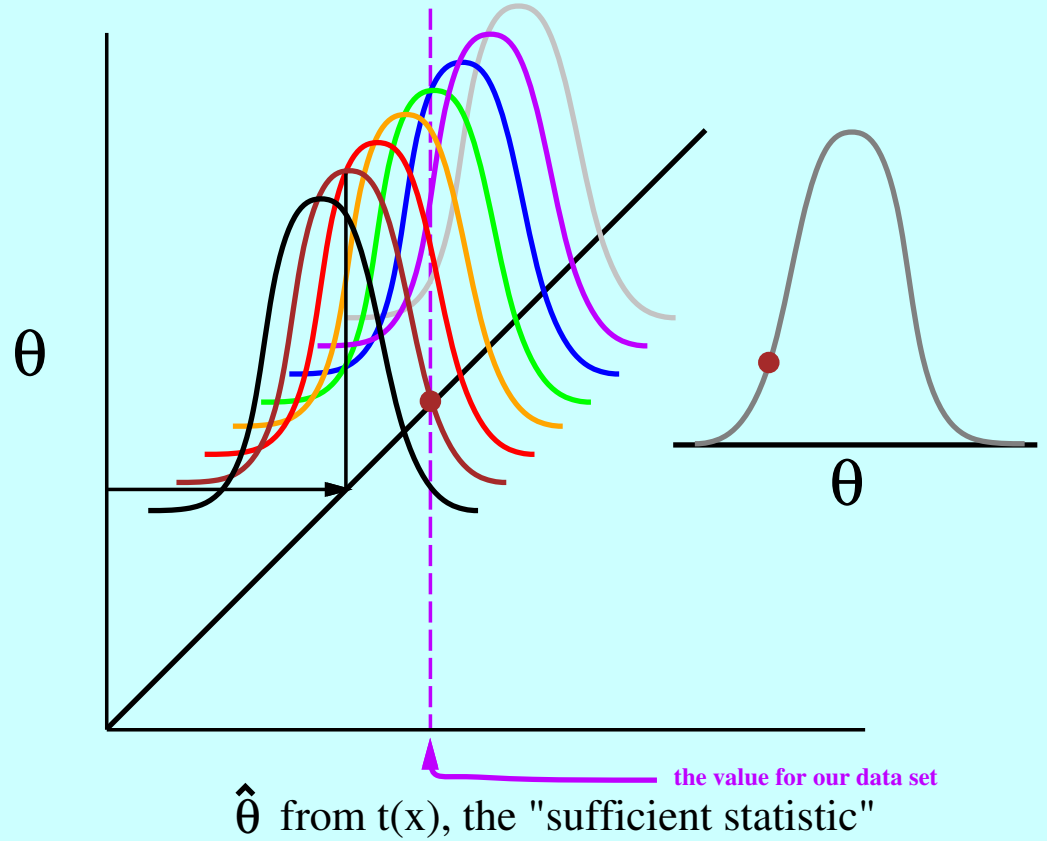
for large amounts of data



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

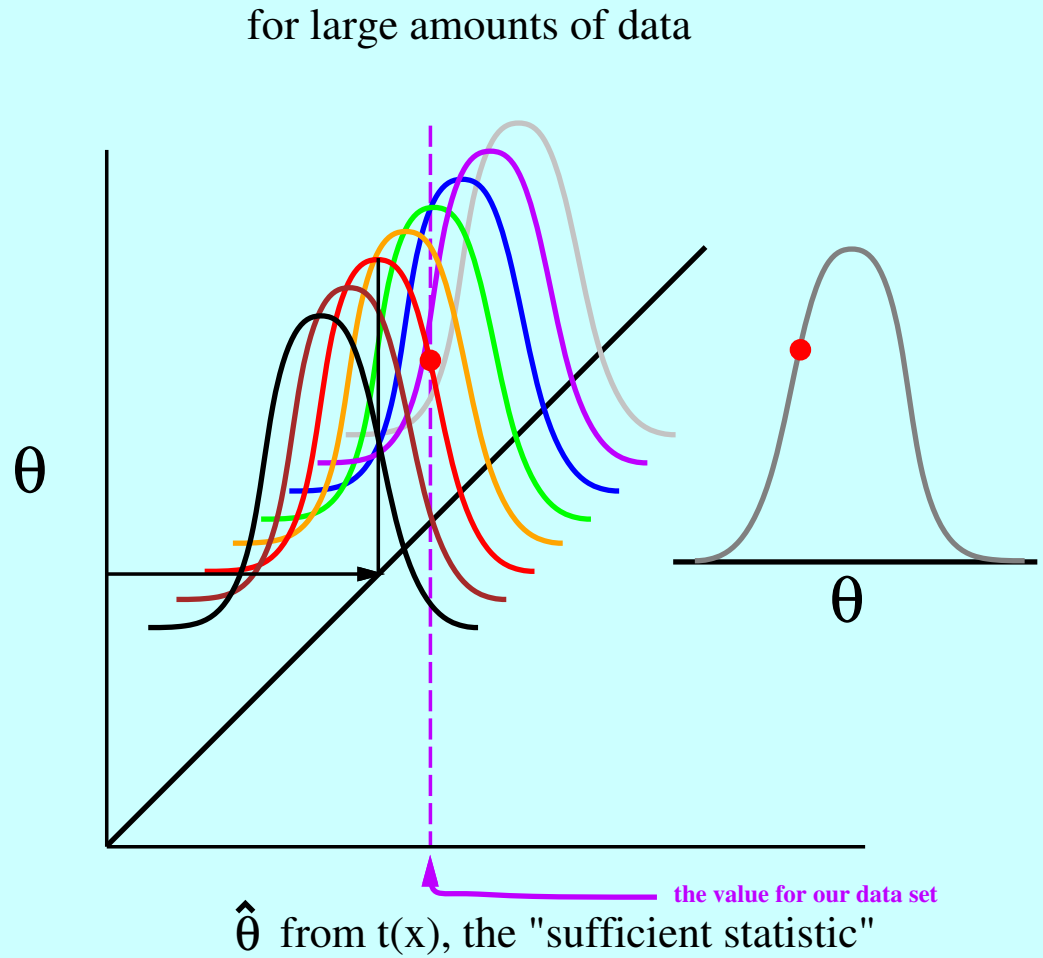
The likelihood curve is nearly a normal distribution

for large amounts of data



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

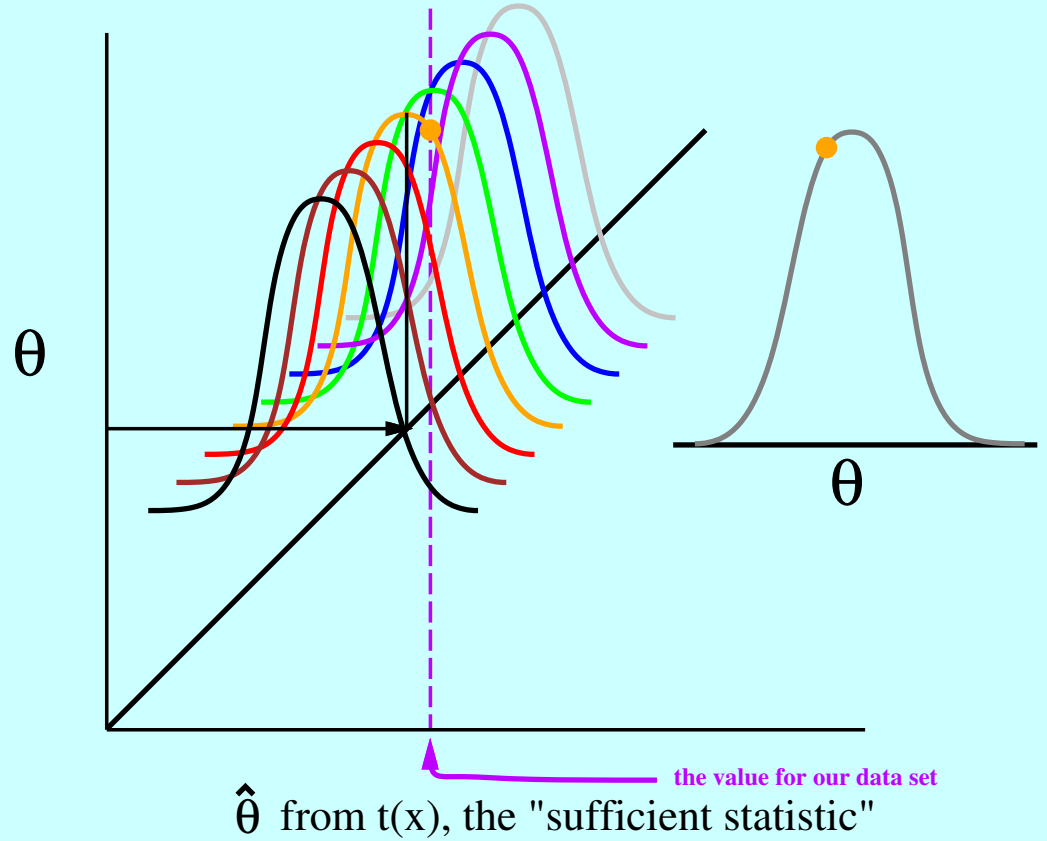
The likelihood curve is nearly a normal distribution



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

The likelihood curve is nearly a normal distribution

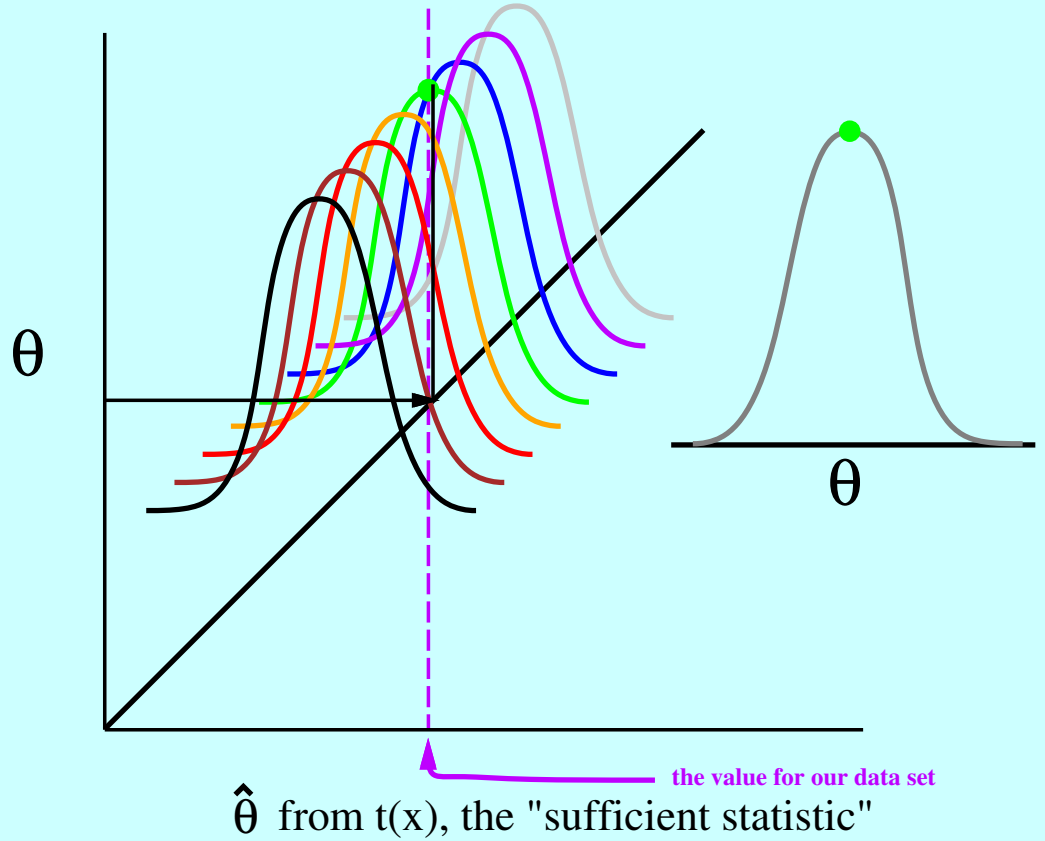
for large amounts of data



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

The likelihood curve is nearly a normal distribution

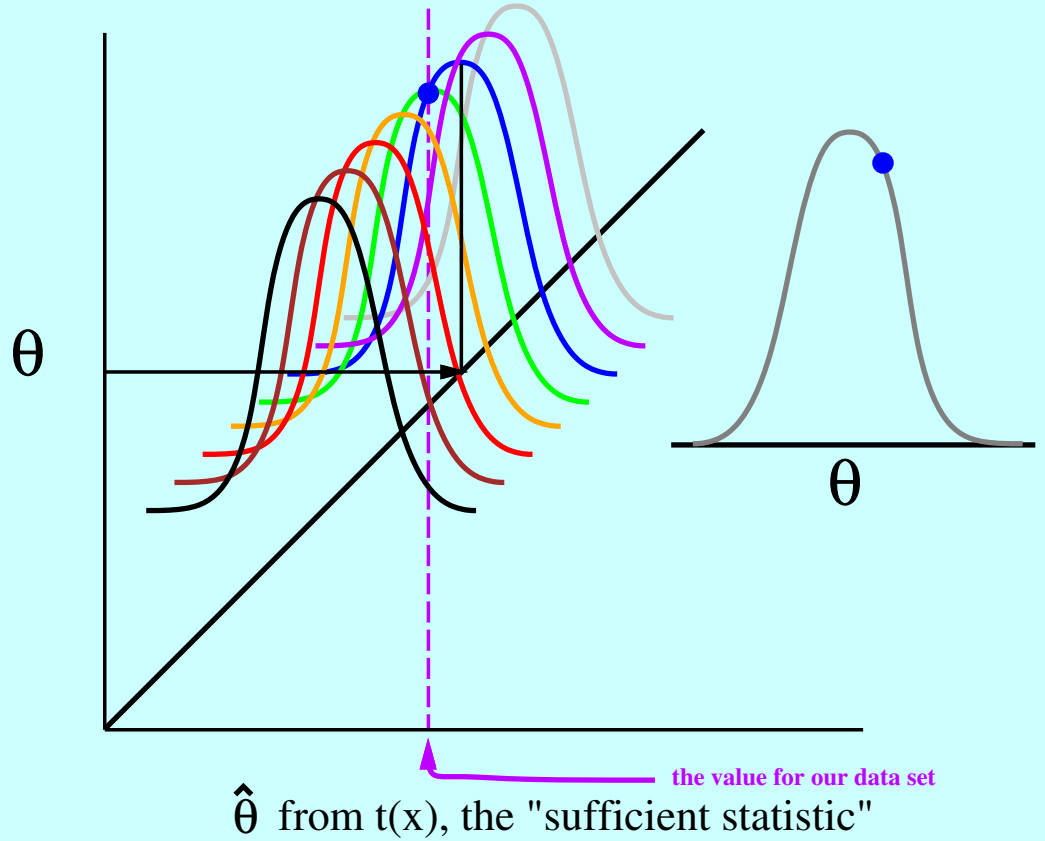
for large amounts of data



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

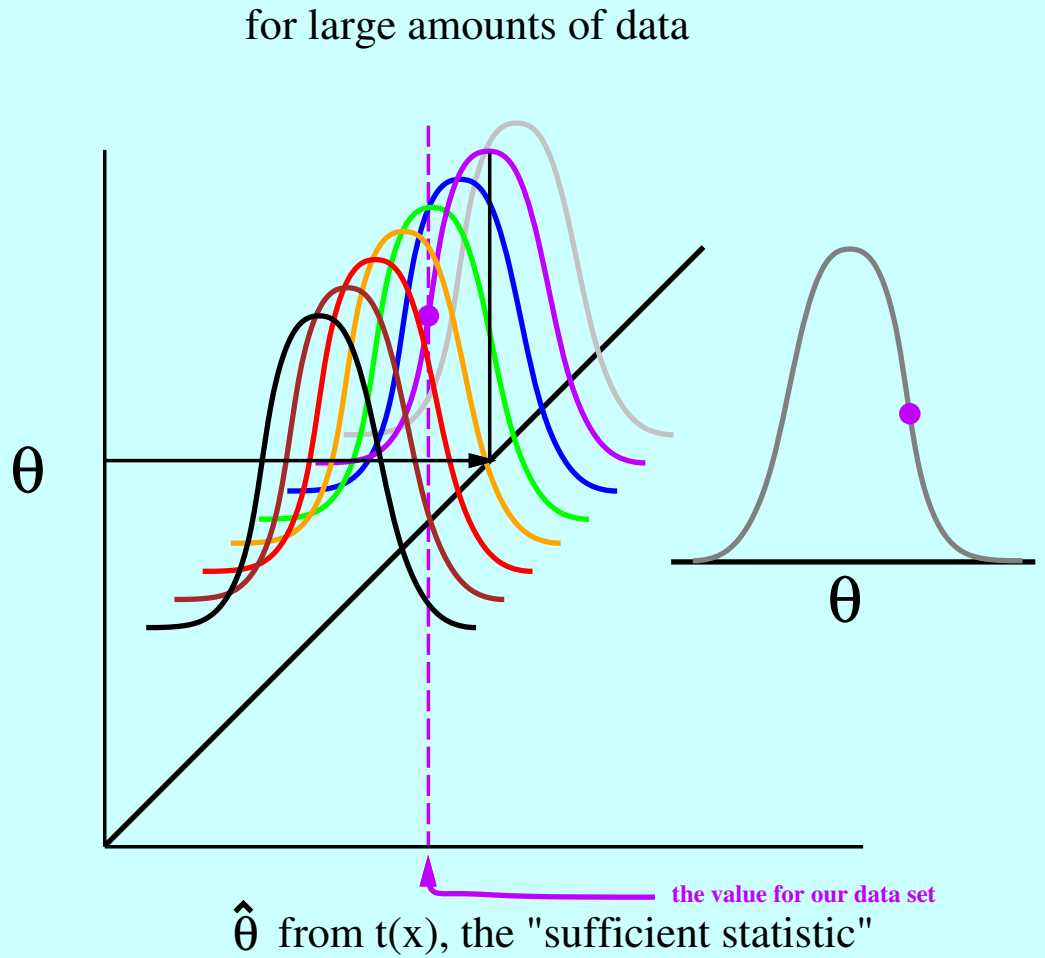
The likelihood curve is nearly a normal distribution

for large amounts of data



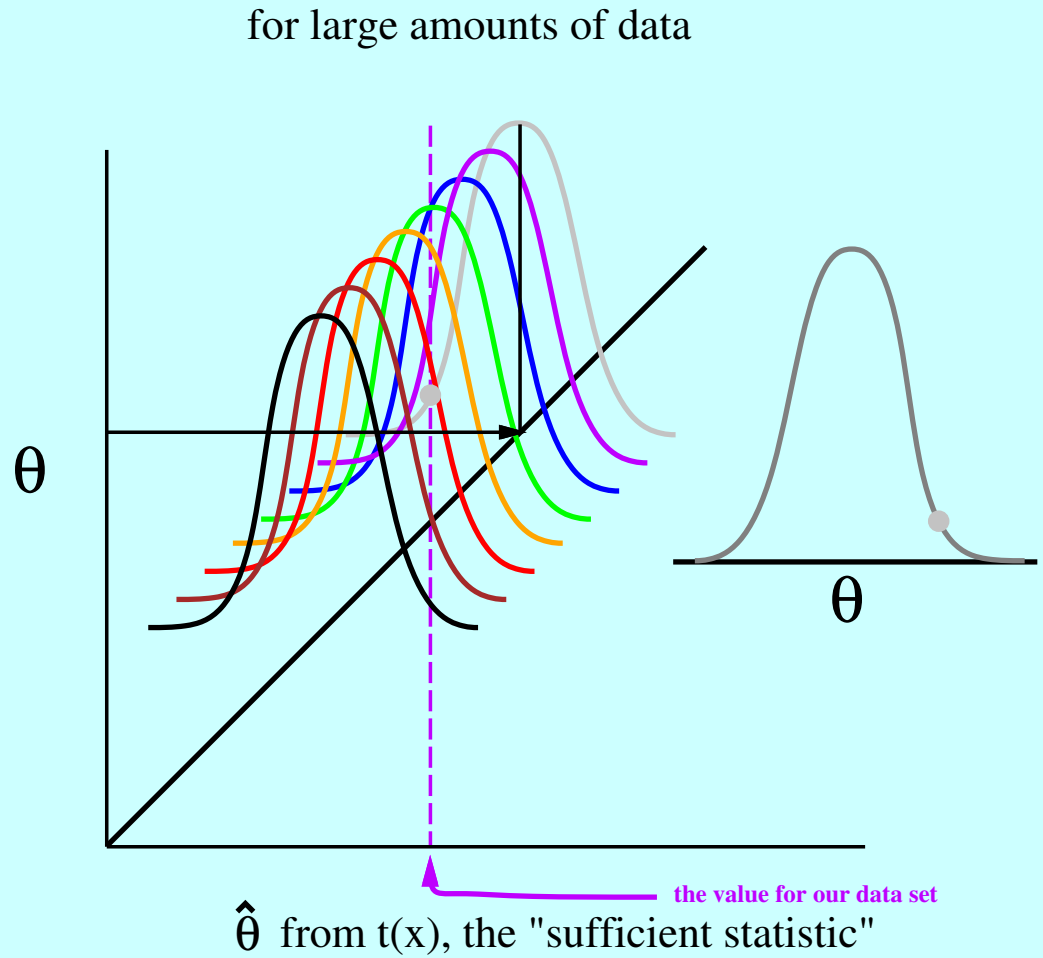
If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

The likelihood curve is nearly a normal distribution



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

The likelihood curve is nearly a normal distribution



If we have large amounts of data, the values of parameters we need to try are all very similar, and the shape of the distribution (which is nearly normal) will not be too different for these values.

Curvatures and covariances of ML estimates

ML estimates have covariances computable from curvatures of the expected log-likelihood:

$$\mathrm{Var}\left[\widehat{\theta}\right] \simeq -1/\left(\frac{\mathsf{d}^2\mathsf{E}(\log(\mathsf{L}))}{\mathsf{d}\theta^2}\right)$$

The same is true when there are multiple parameters:

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}\right] \simeq \mathbf{V} \simeq -\mathbf{C}^{-1}$$

where

$$\mathsf{C}_{\mathsf{ij}} = \mathsf{E}\left(\frac{\partial^2 \log(\mathsf{L})}{\partial\theta_{\mathsf{i}}\,\partial\theta_{\mathsf{j}}}\right)$$

With large amounts of data, asymptotically

When the true value of θ is θ_0,

$$\frac{\hat{\theta} - \theta_0}{\sqrt{v}} \sim \mathcal{N}(0, 1)$$

Since $1/v$ is the negative of the curvature of the log-likelihood:

$$\ln L(\theta_0) = \ln L(\hat{\theta}) - \frac{1}{2}\frac{(\theta_0 - \hat{\theta})^2}{v}$$

so that twice the difference of log-likelihoods is the square of a normal:

$$2\left(\ln L(\hat{\theta}) - \ln L(\theta_0)\right) \sim \chi^2_1$$

Corresponding results for multiple parameters

$$\ln \mathsf{L}(\theta_0) \simeq \ln \mathsf{L}(\theta_0) - \frac{1}{2} (\theta_0 - \theta)^{\mathsf{T}} \mathbf{C} (\theta_0 - \theta)$$

$$(\theta - \theta_0)^{\mathsf{T}} \mathbf{C} (\theta - \theta_0) \sim \chi^2_{\mathsf{p}}$$

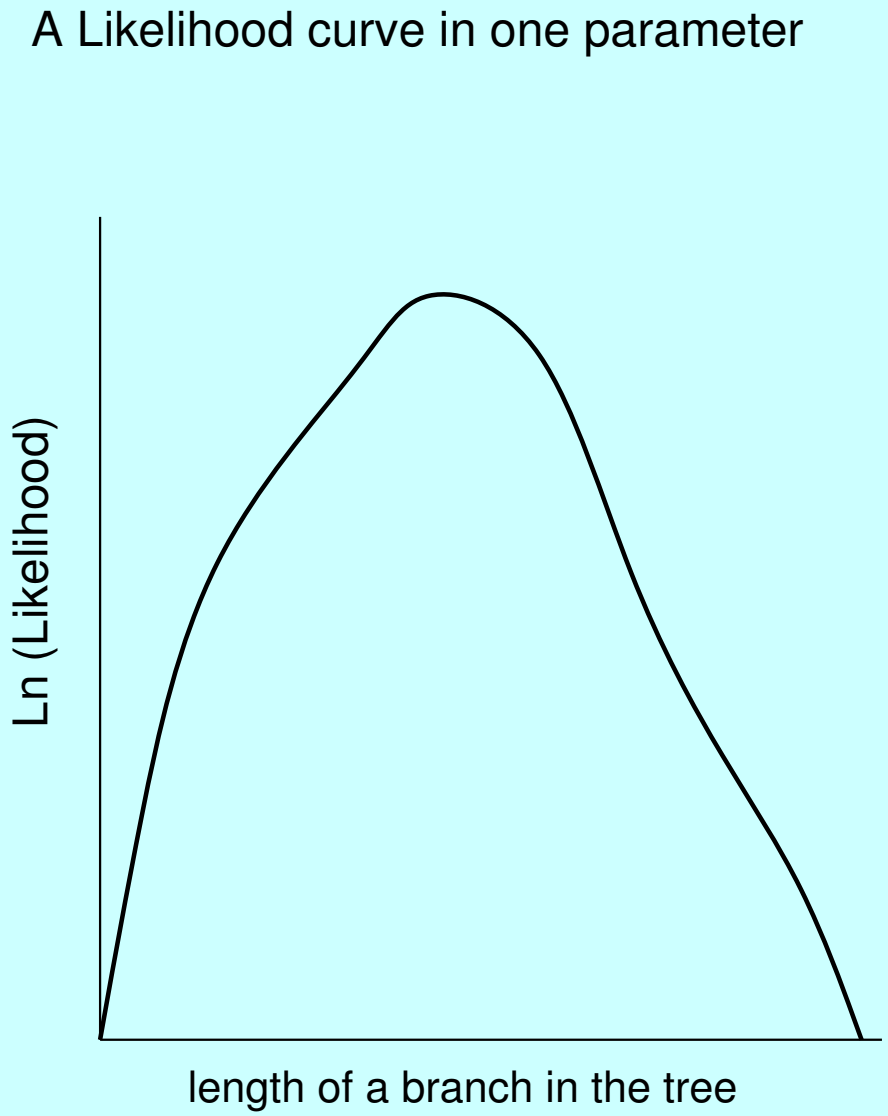
so that the log-likelihood difference is:

$$2 \left(\ln \mathsf{L}(\hat{\theta}) - \ln \mathsf{L}(\theta_0) \right) \sim \chi^2_{\mathsf{p}}$$

When in the (true) null hypothesis θ_0 we have q of the p parameters constrained:

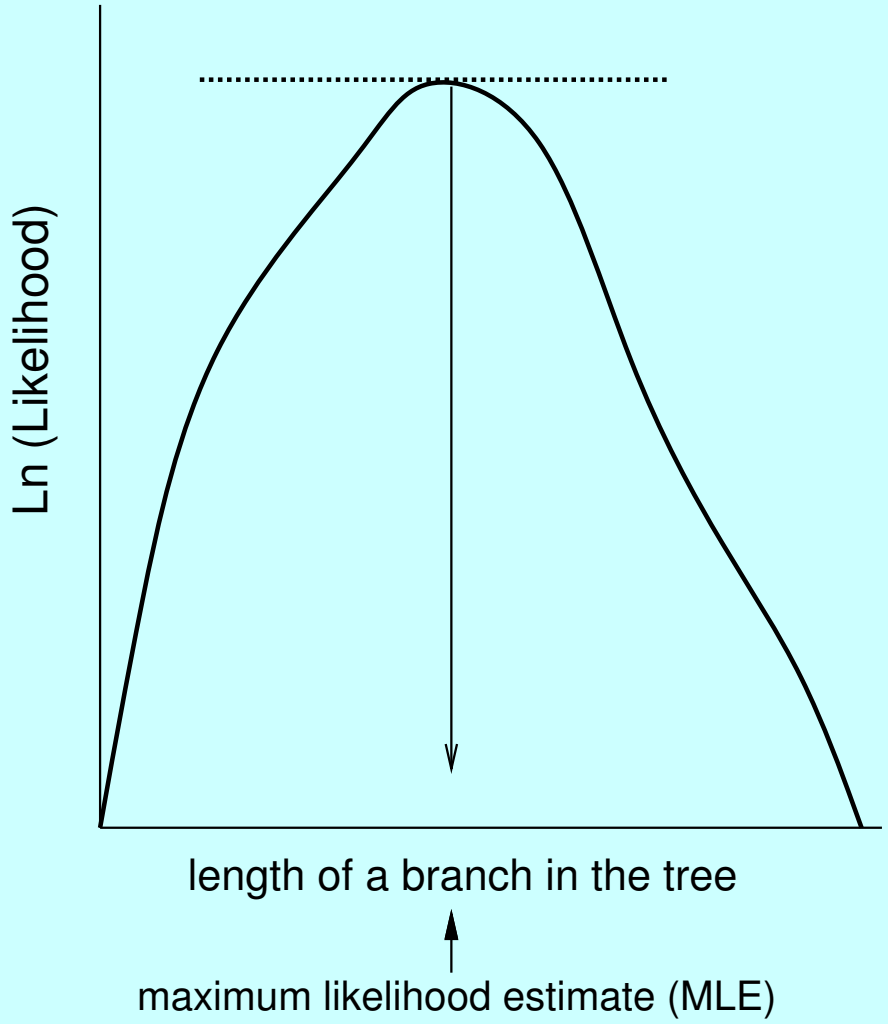
$$2 \left(\ln \mathsf{L}(\hat{\theta}) - \ln \mathsf{L}(\theta_0) \right) \sim \chi^2_{\mathsf{q}}$$

A log-likelihood curve



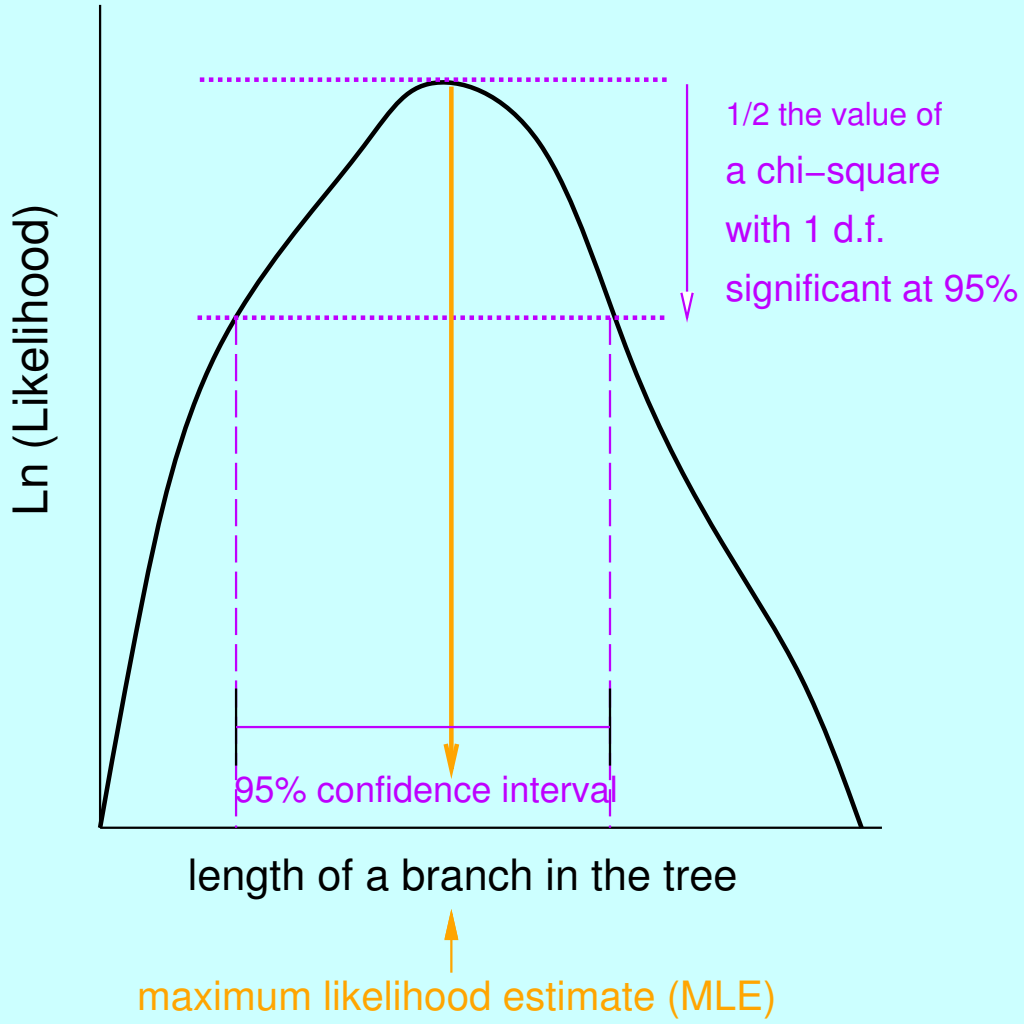
Its maximum likelihood estimate

A Likelihood curve in one parameter
and the maximum likelihood estimate

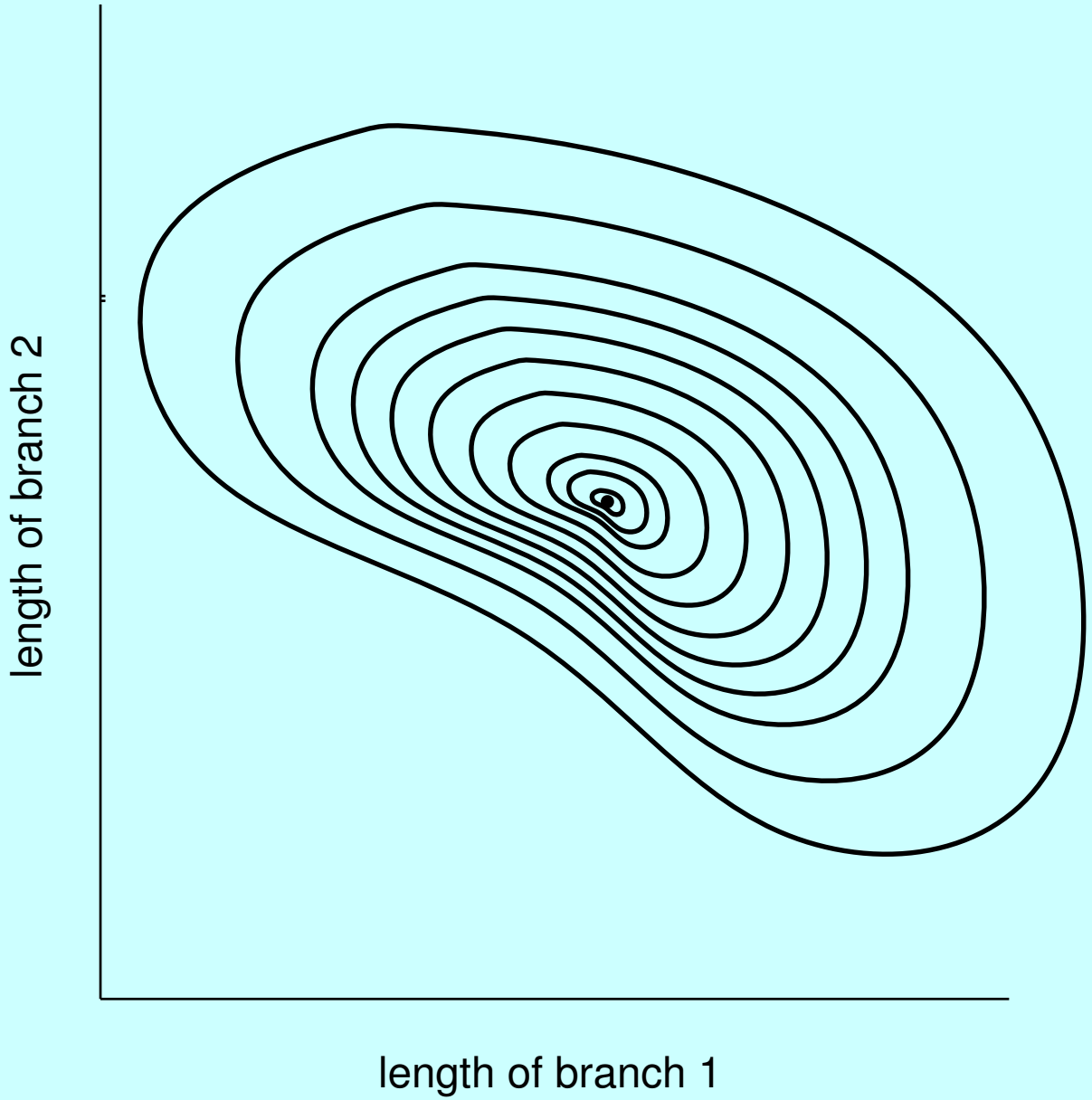


The (approximate, asymptotic) confidence interval

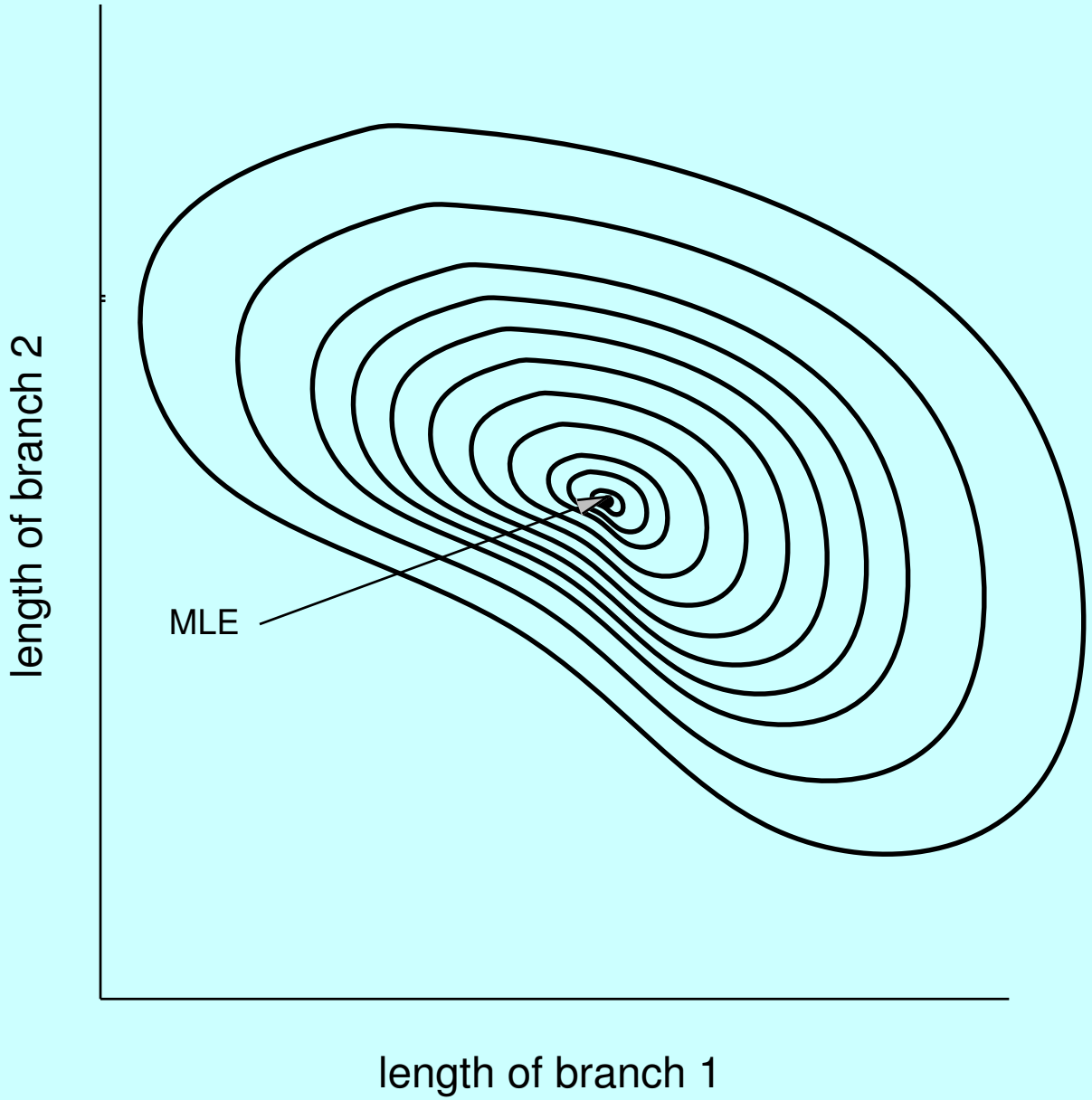
A Likelihood curve in one parameter
and the maximum likelihood estimate and
confidence interval derived from it



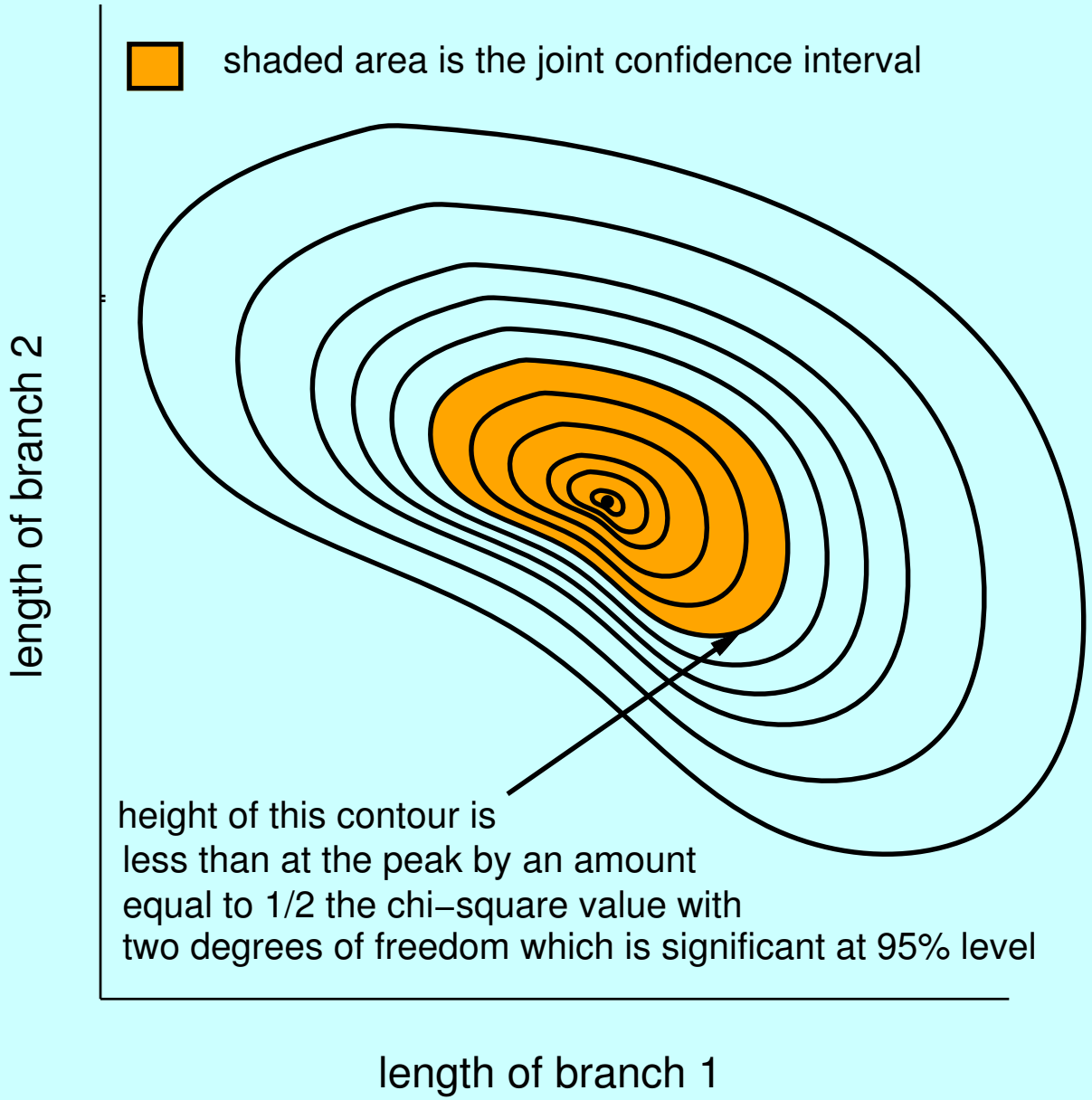
Contours of a log-likelihood surface in two dimensions



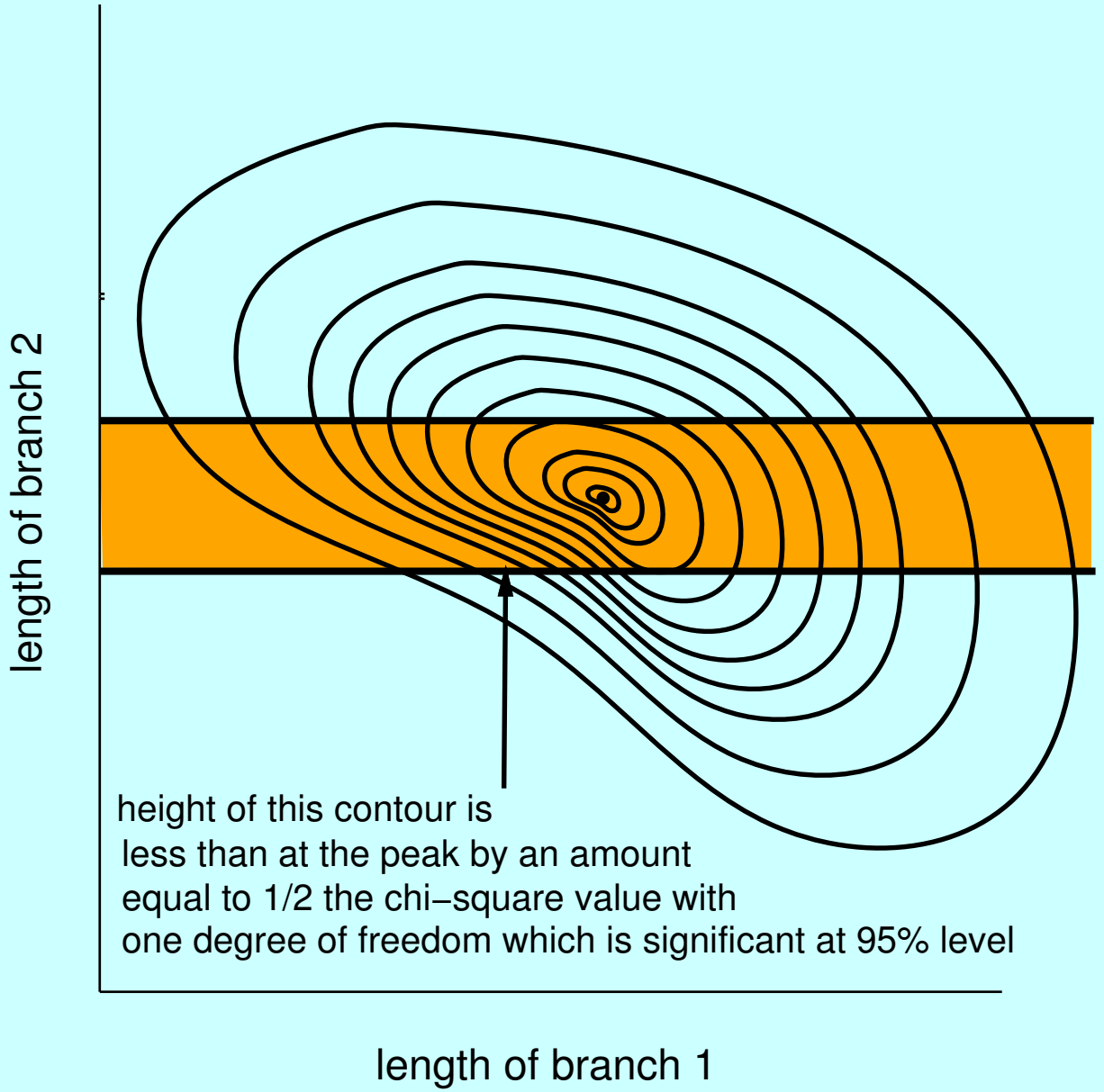
Contours of a log-likelihood surface in two dimensions



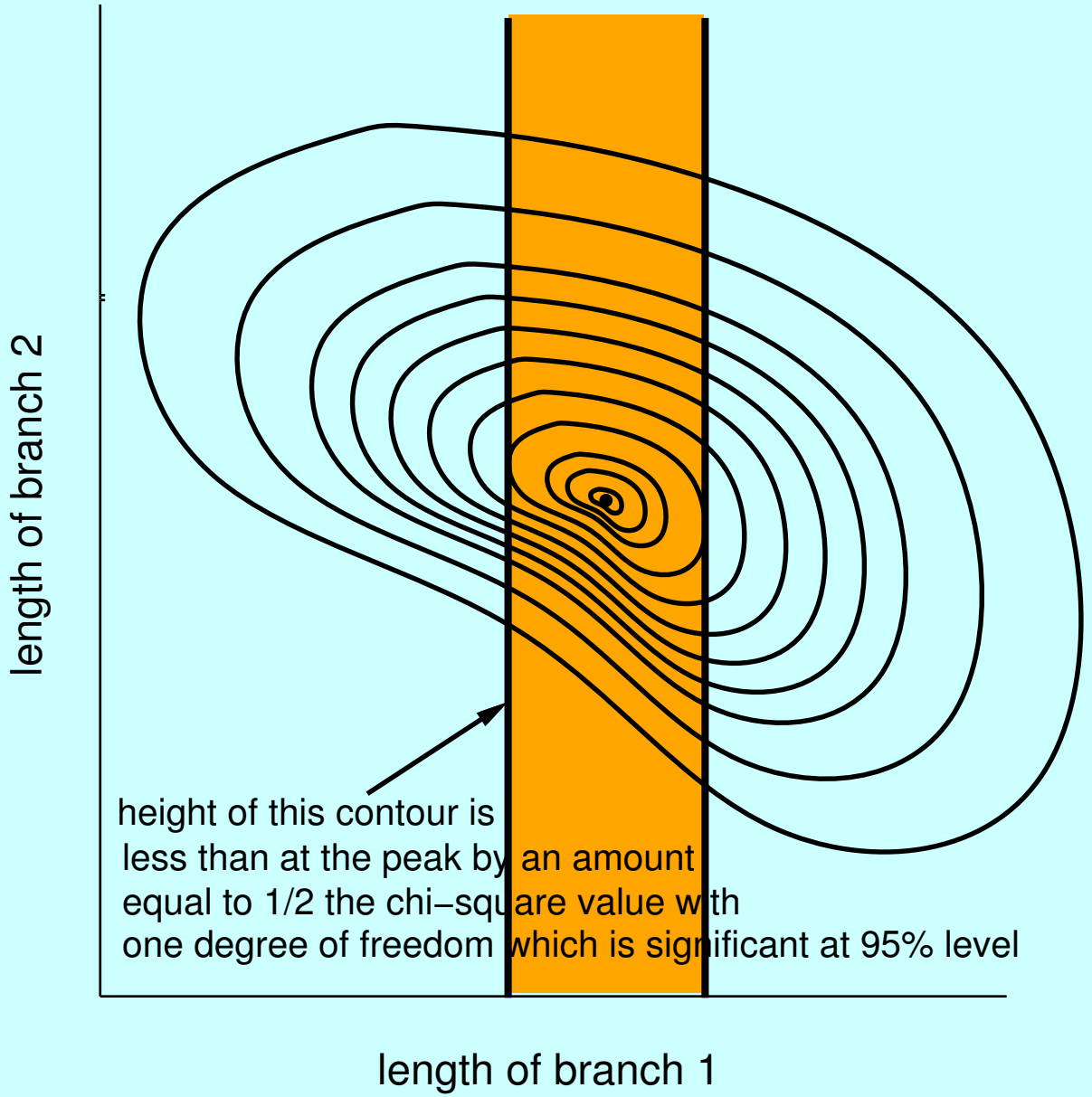
Log-likelihood-based confidence set for two variables



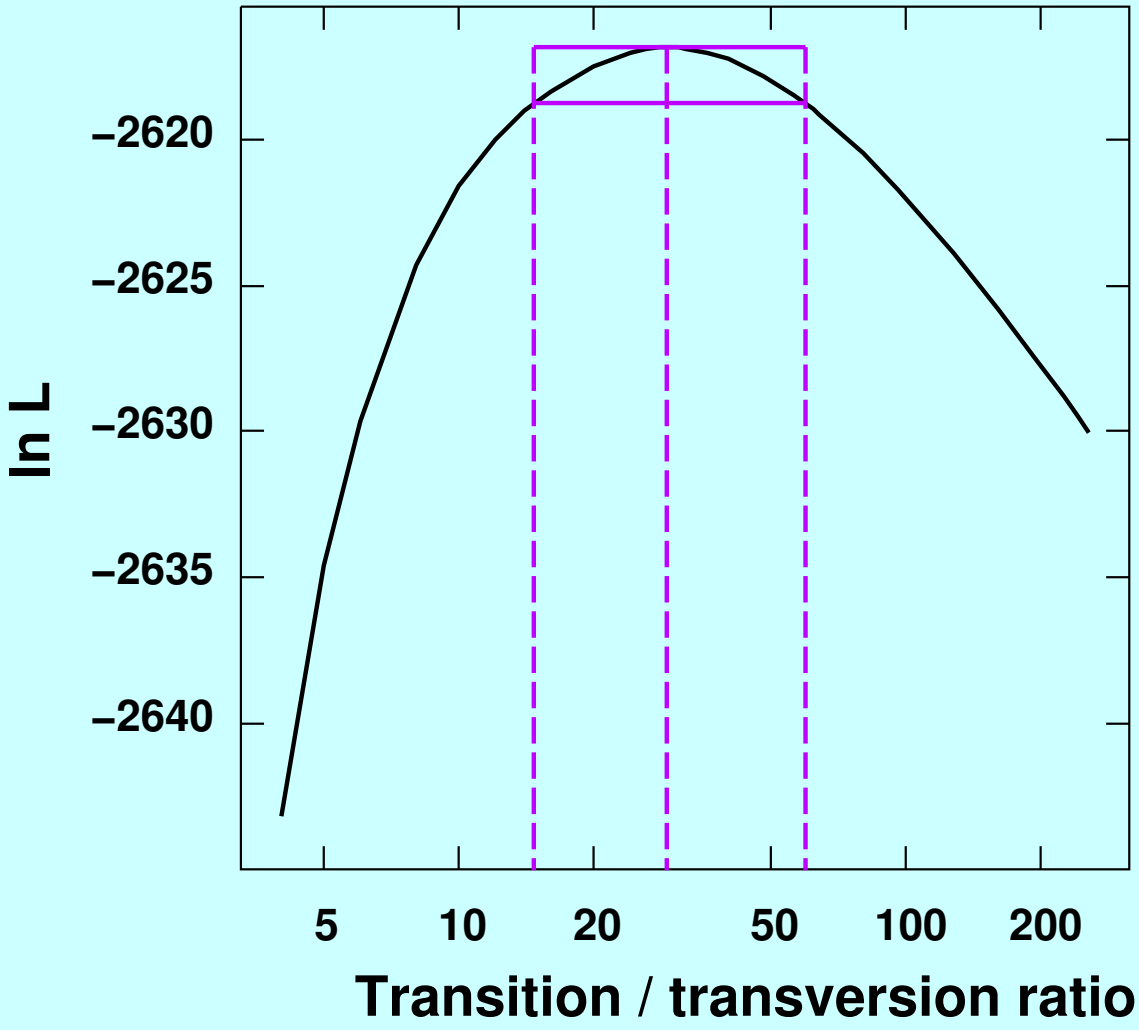
Confidence interval for one variable



Confidence interval for the other variable

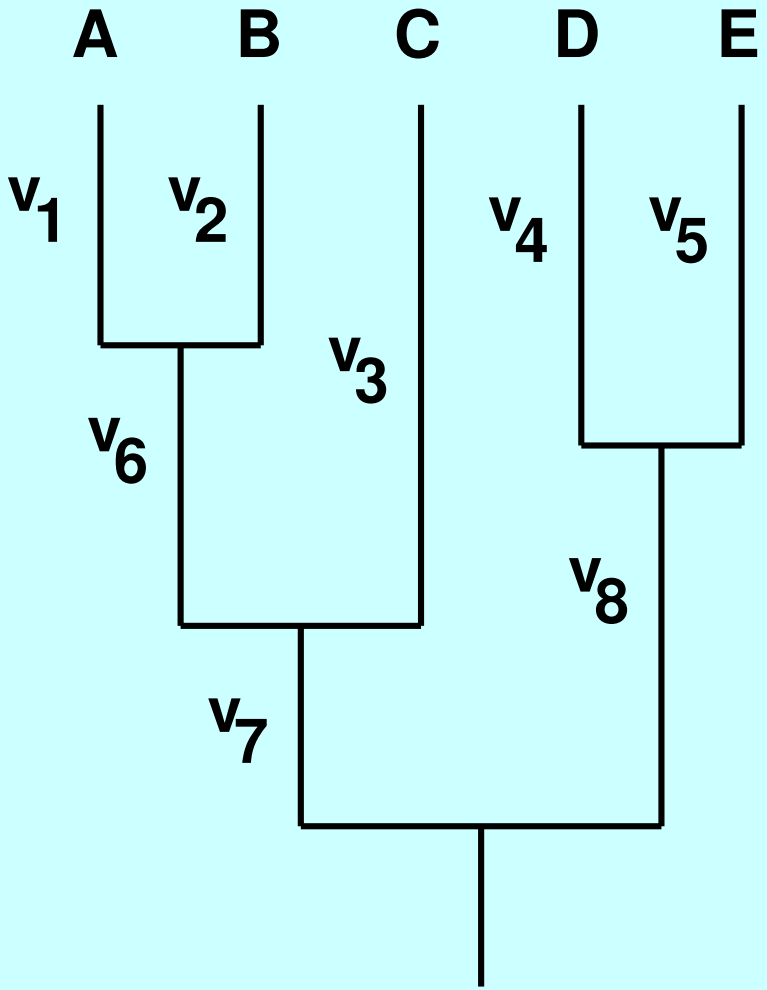


Likelihood ratio interval for a parameter



Inferring the transition/transversion ratio for an F84 model with the 14-species primate mitochondria data set.

LRT of a molecular clock – how many parameters?



Constraints for a clock

$$v_1 = v_2$$

$$v_4 = v_5$$

$$v_1 + v_6 = v_3$$

$$v_3 + v_7 = v_4 + v_8$$

How does each equation constrain the branch lengths in the unrooted tree? What about the red equation?

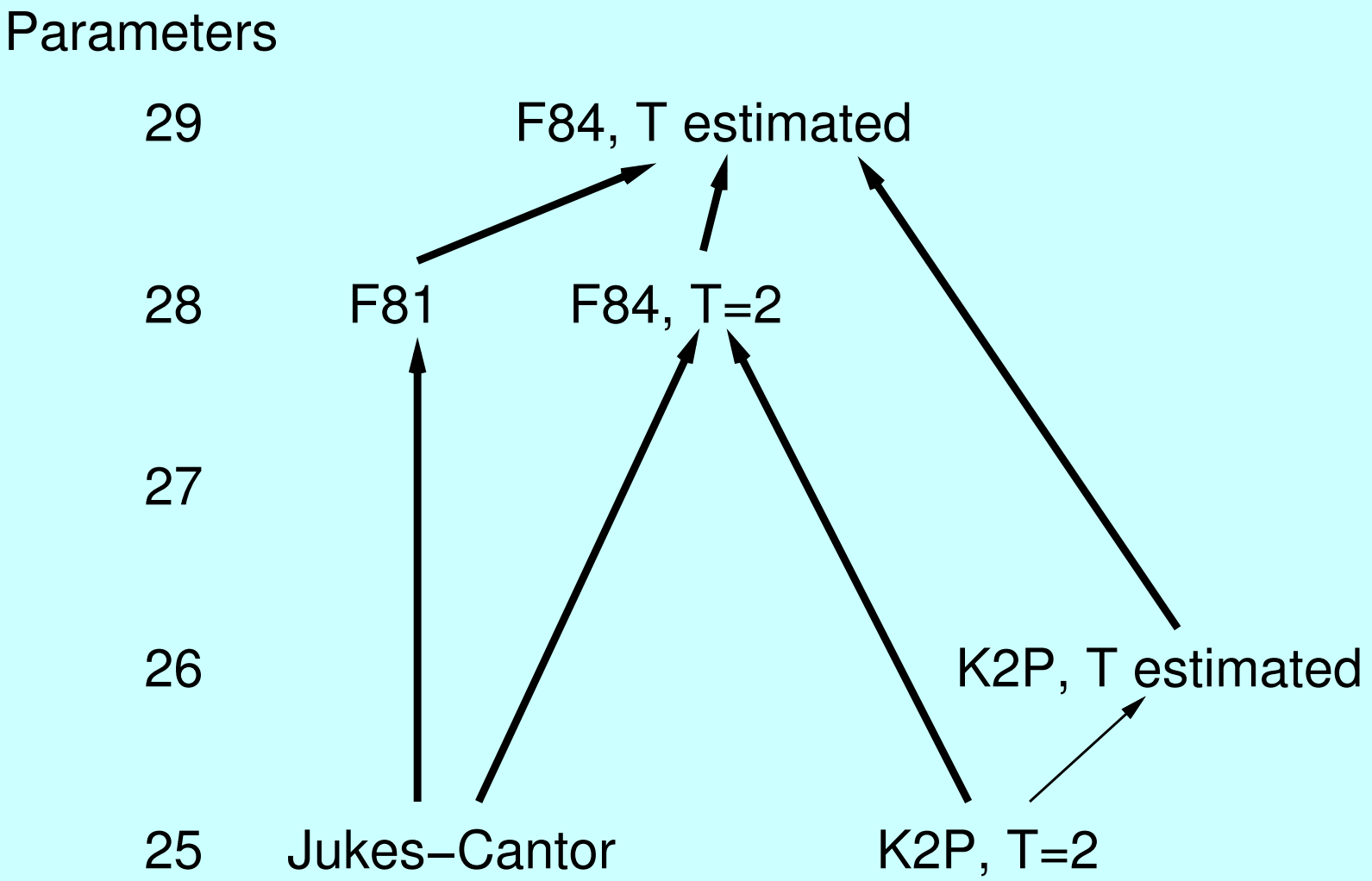
Likelihood Ratio Test for a molecular clock

Using the 7-species mitochondrial DNA data set (the great apes plus Bovine and Mouse), we get with Ts/Tn = 30 and an F84 model:

Tree	$\ln$ L
No clock	-1372.77620
Clock	-1414.45053
Difference	41.67473

Chi-square statistic: $2 \times 41.675 = 83.35$, with $n - 2 = 5$ degrees of freedom – highly significant.

Model selection using the LRT



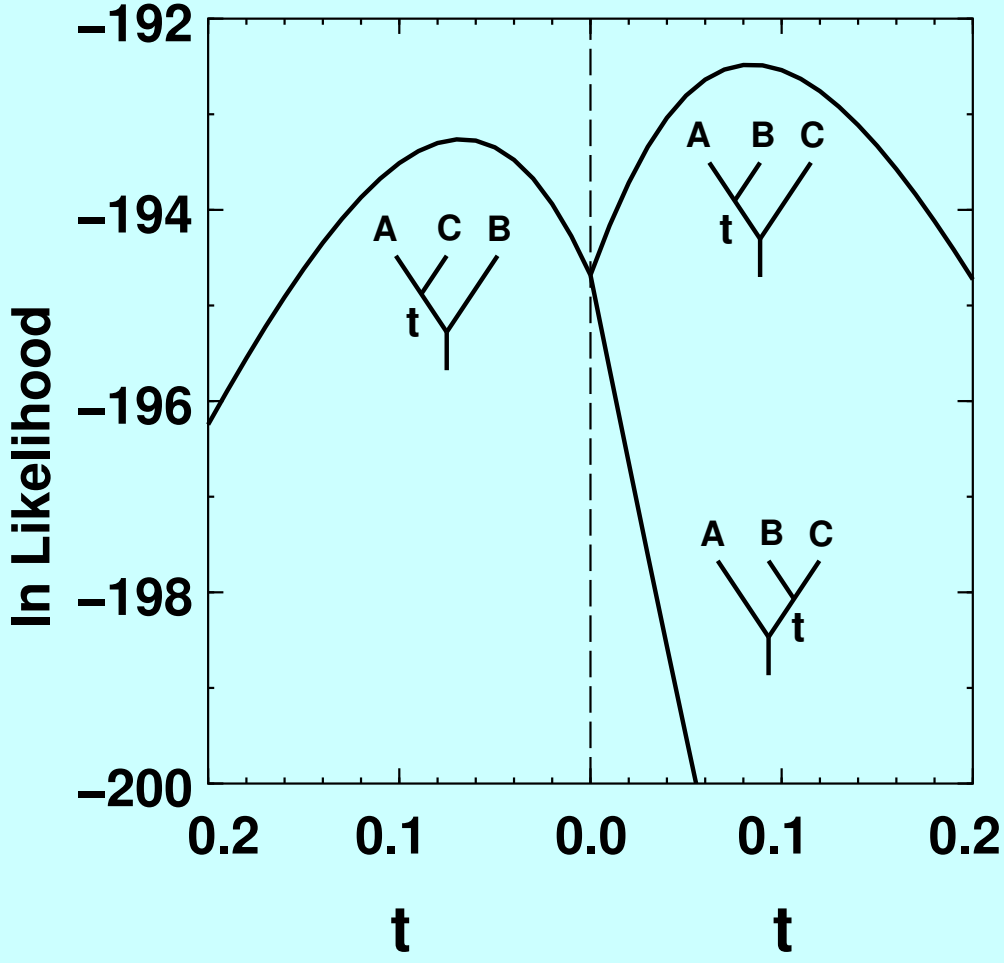
The problem with using likelihood ratio tests is the multiplicity of tests and the multiple routes to the same hypotheses.

The Akaike Information Criterion

Compare between hypotheses $-2\ln \mathsf{L} + 2\mathsf{p}$ (the same as reducing the log-likelihood by the number of parameters)

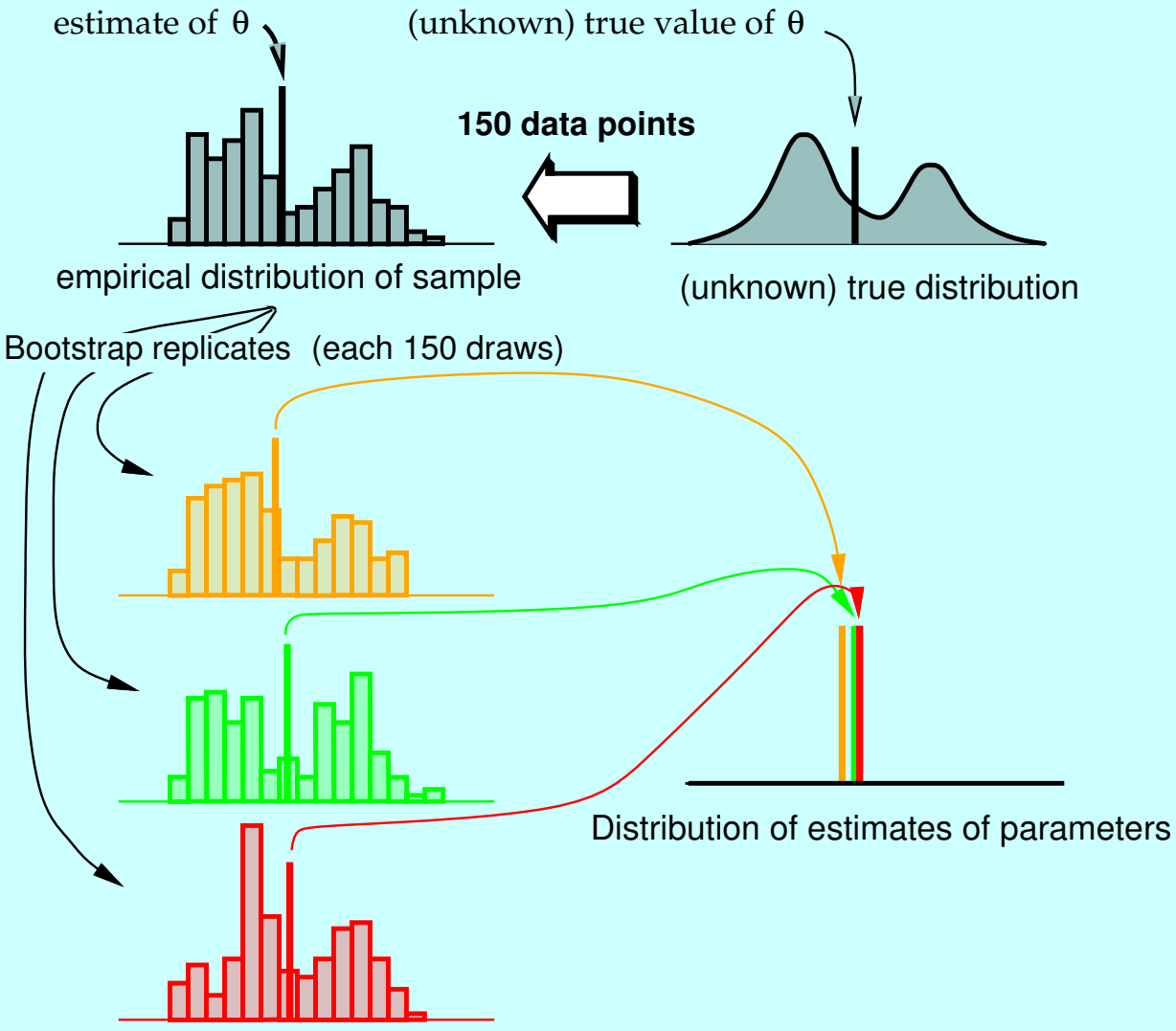
Model	$\ln \mathsf{L}$	Number of parameters	AIC
Jukes-Cantor	-3068.29186	25	6186.58
K2P, $\mathsf{R} = 2.0$	-2953.15830	25	5956.32
K2P, $\widehat{\mathsf{R}} = 1.889$	-2952.94264	26	5957.89
F81	-2935.25430	28	5926.51
F84, $\mathsf{R} = 2.0$	-2680.32982	28	5416.66
F84, $\widehat{\mathsf{R}} = 28.95$	-2616.3981	29	5290.80

Can we test trees using the LRT?



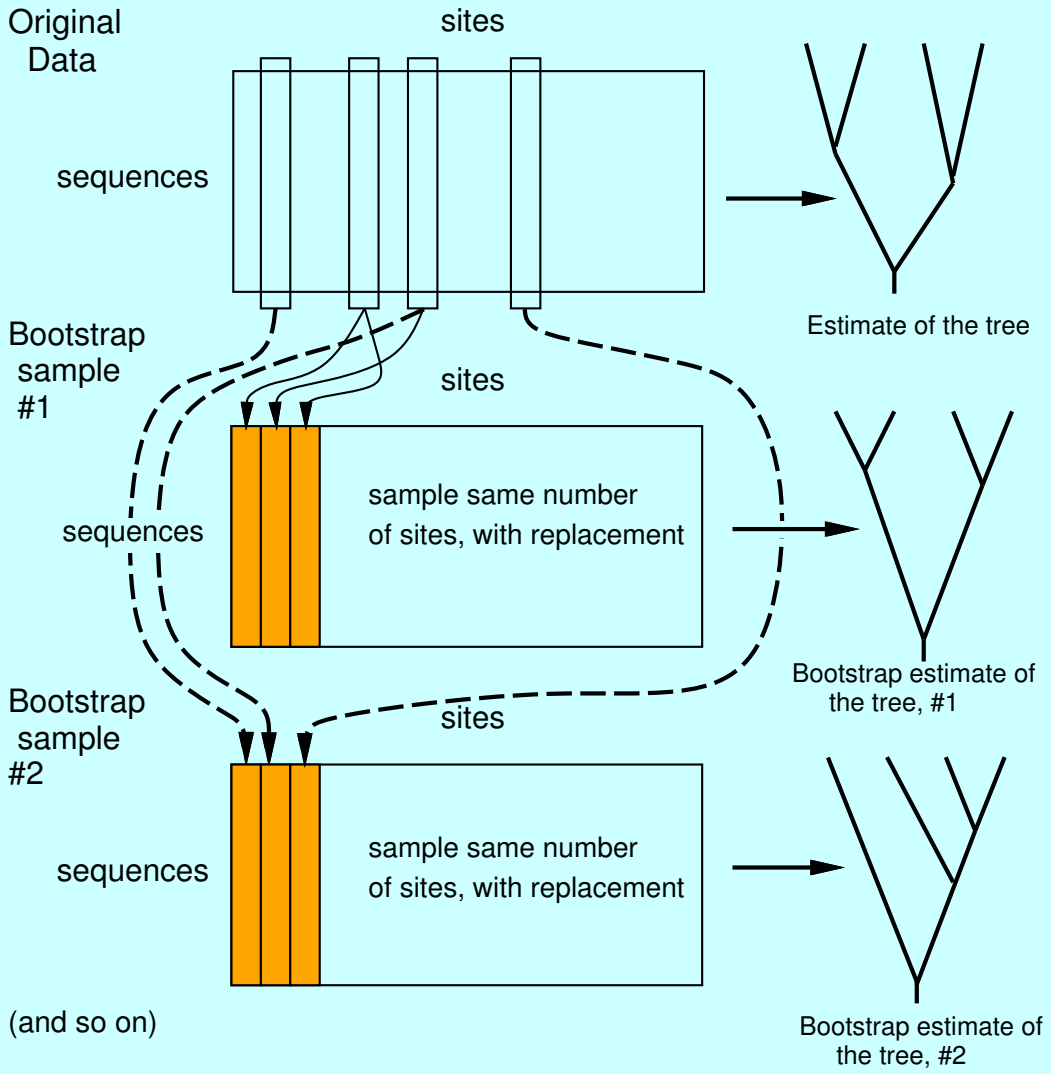
If so, how many degrees of freedom for the comparison of the two peaks? These are three-species clocklike trees (shown here plotted in a “profile log-likelihood plot” plotting the highest likelihood for each value of the interior branch length).

The bootstrap



An example with mixed normal distributions. Draw from the empirical distribution 150 times if there are 150 data points. With replacement!

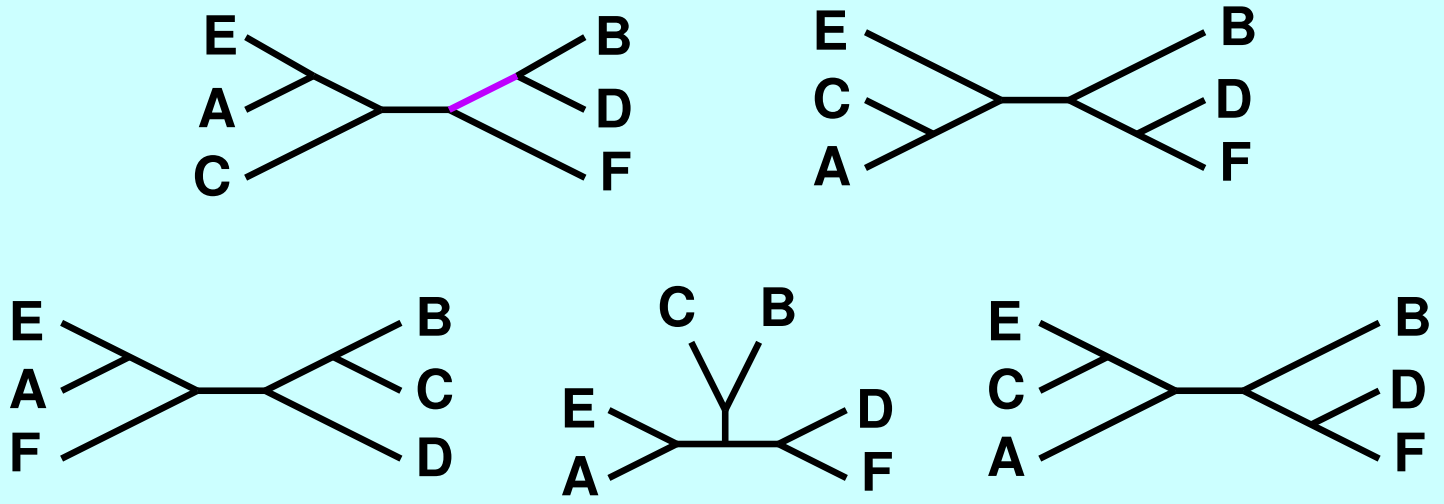
The bootstrap for phylogenies



Drawing columns of the data matrix, with replacement.

A partition defined by a branch in the first tree

Trees:

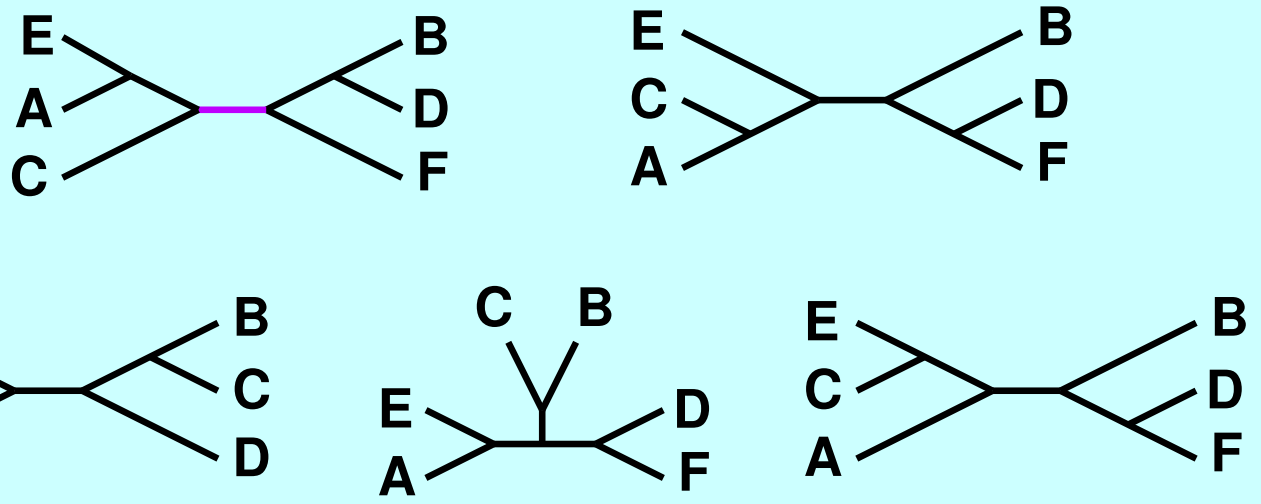


How many times each partition of species is found:

AE | BCDF
ACE | BDF
ACEF | BD **1**
AC | BDEF
AEF | BCD
ADEF | BC
ABDF | EC
ABCE | DF

Another partition from the first tree

Trees:

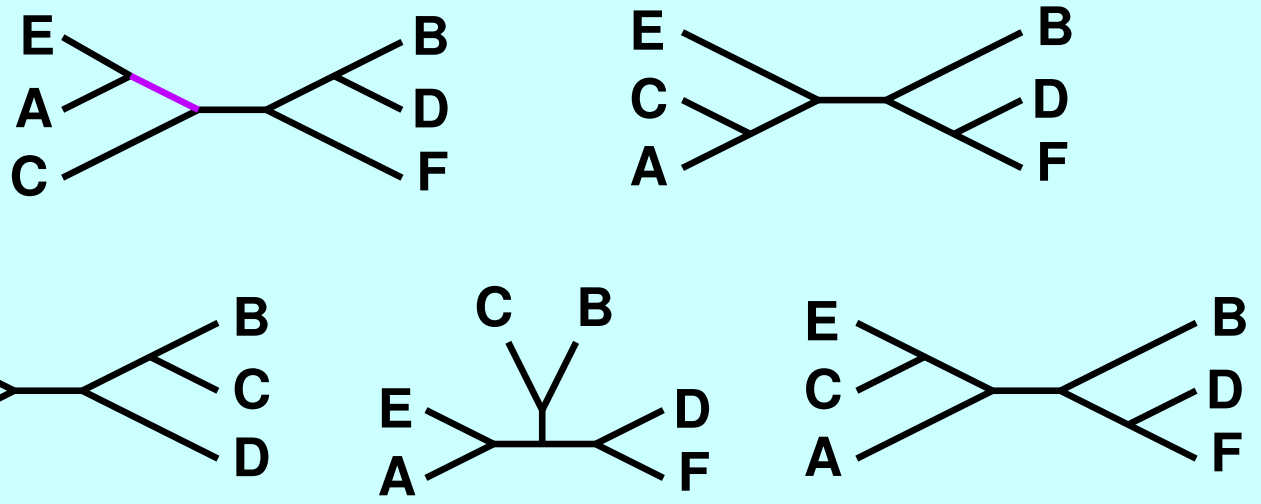


How many times each partition of species is found:

AE | BCDF	
ACE | BDF	**1**
ACEF | BD	**1**
AC | BDEF	
AEF | BCD	
ADEF | BC	
ABDF | EC	
ABCE | DF	

The third partition from that tree

Trees:

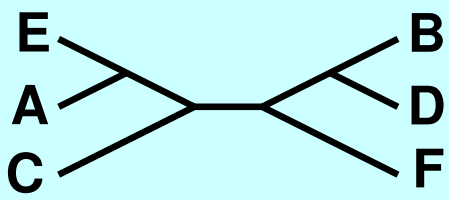


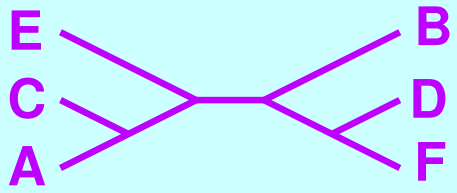
How many times each partition of species is found:

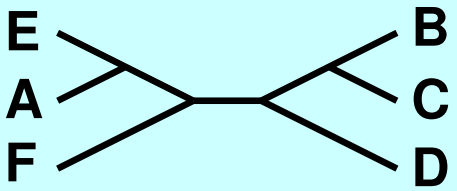
AE | BCDF	**1**
ACE | BDF	**1**
ACEF | BD	**1**
AC | BDEF	
AEF | BCD	
ADEF | BC	
ABDF | EC	
ABCE | DF	

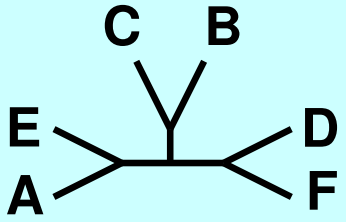
Partitions from the second tree

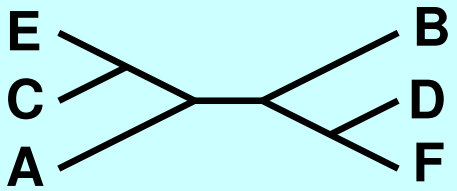
Trees:









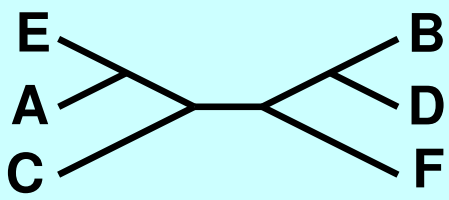


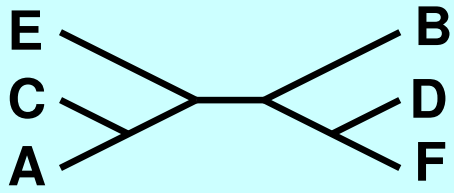
How many times each partition of species is found:

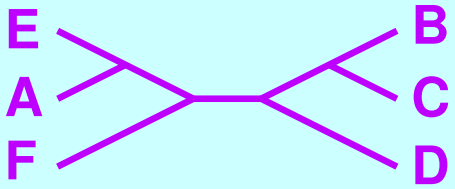
AE | BCDF	**1**
ACE | BDF	**2**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	
ADEF | BC	
ABDF | EC	
ABCE | DF	**1**

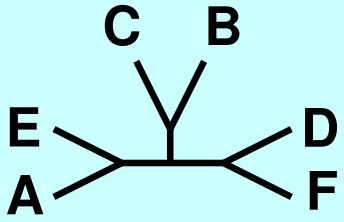
Partitions from the third tree

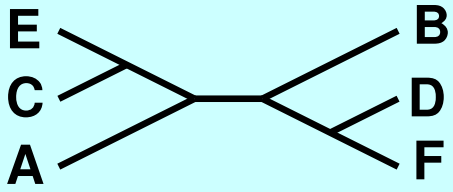
Trees:









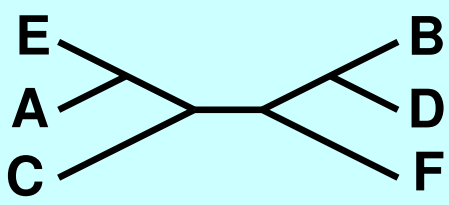


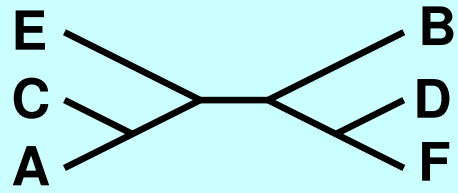
How many times each partition of species is found:

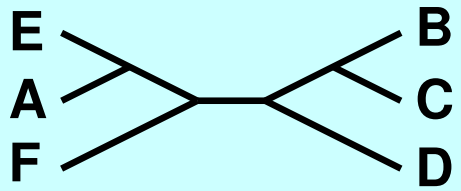
AE | BCDF	**2**
ACE | BDF	**2**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**1**
ABDF | EC	
ABCE | DF	**1**

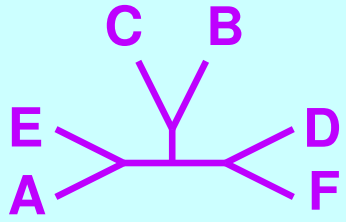
Partitions from the fourth tree

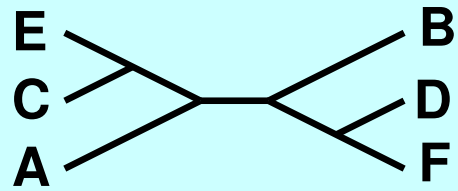
Trees:









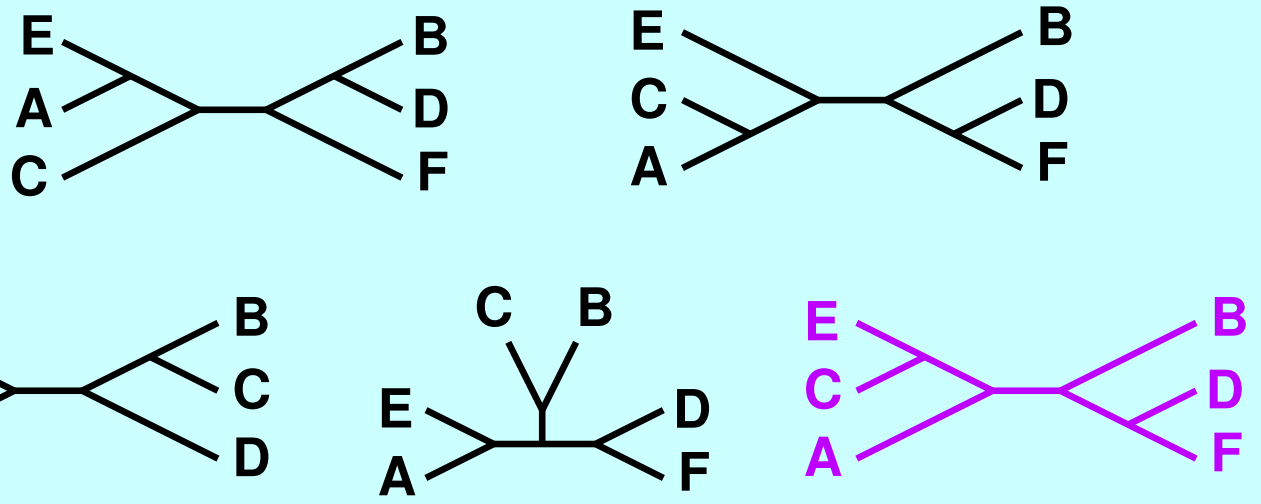


How many times each partition of species is found:

AE | BCDF	**3**
ACE | BDF	**2**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	
ABCE | DF	**2**

Partitions from the fifth tree

Trees:

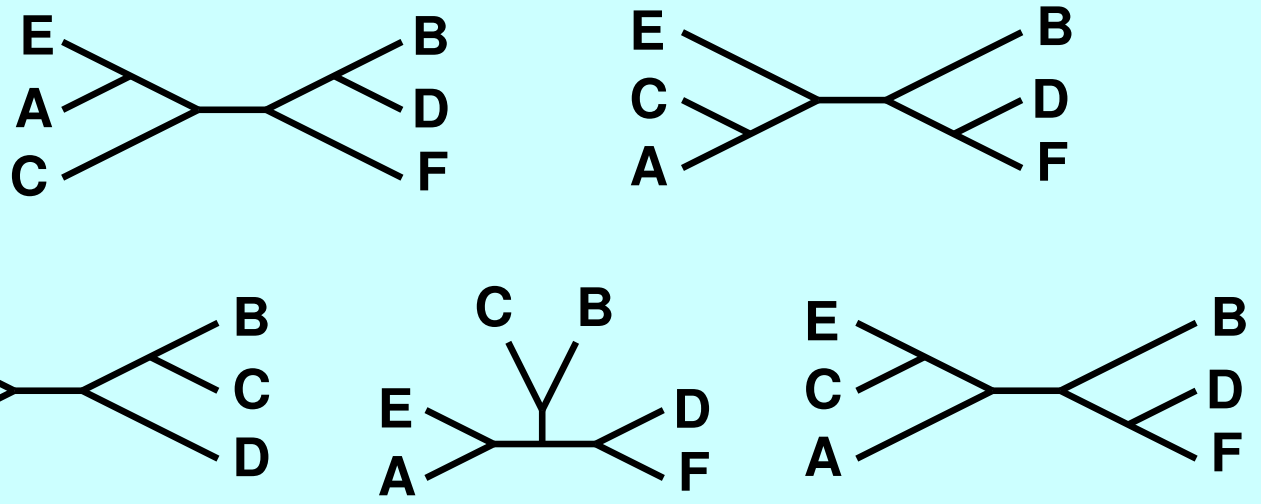


How many times each partition of species is found:

AE | BCDF	**3**
ACE | BDF	**3**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	**1**
ABCE | DF	**3**

The table of partitions from all trees

Trees:

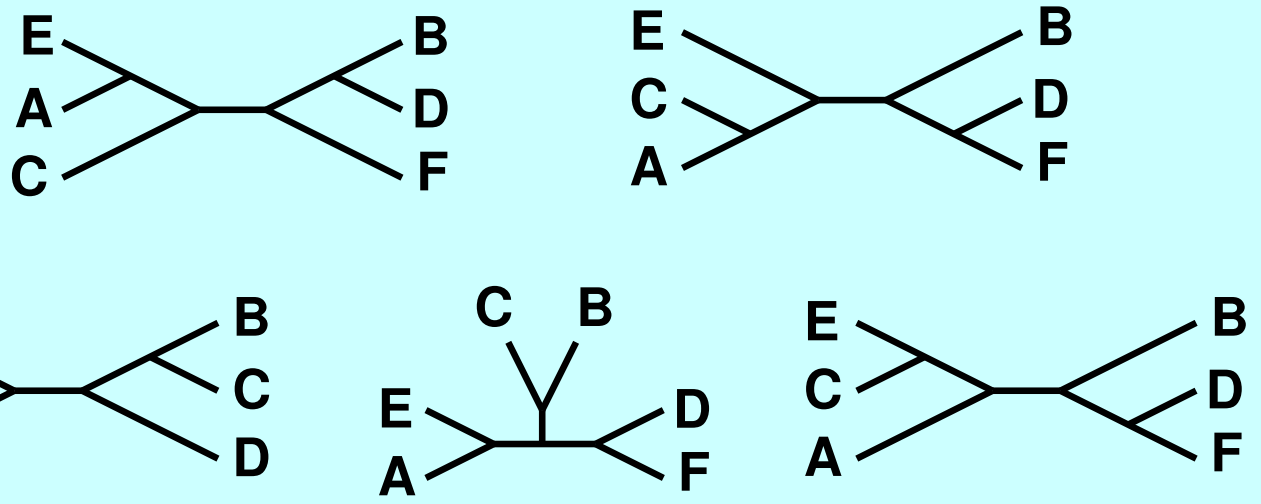


How many times each partition of species is found:

AE | BCDF	**3**
ACE | BDF	**3**
ACEF | BD	**1**
AC | BDEF	**1**
AEF | BCD	**1**
ADEF | BC	**2**
ABDF | EC	**1**
ABCE | DF	**3**

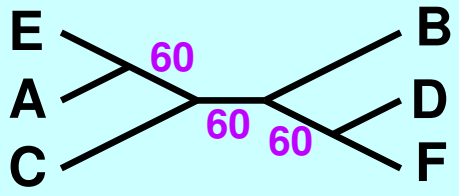
The majority-rule consensus tree

Trees:



How many times each partition of species is found:

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



Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.

Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.

Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.

Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
- Thus the set of these “characters” that occur in more then 50% of the trees are all pairwise compatible.

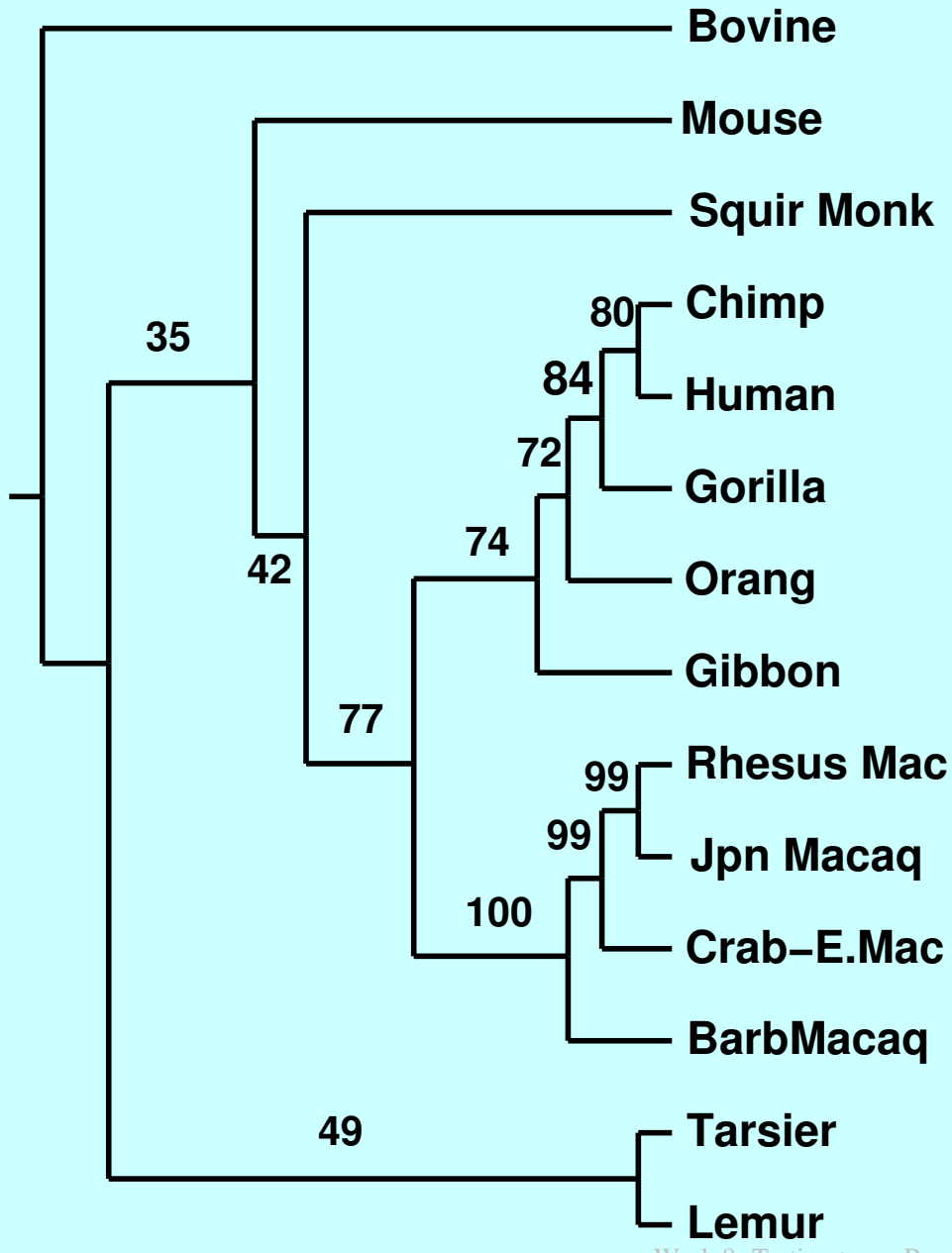
Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
- Thus the set of these "characters" that occur in more then 50% of the trees are all pairwise compatible.
- By the Pairwise Compatibility Theorem (remember that?) they must then be jointly compatible

Why will the MR consensus give a tree?

- Suppose that for each partition in a tree we construct a (fake) morphological character with 0 for one set in the partition, 1 for the other.
- Such a character is compatible with a tree if (and only if) the tree contains that partition.
- If two of these characters both occur in more than 50% of the trees, they must co-occur in at least one tree.
- Thus the set of these "characters" that occur in more then 50% of the trees are all pairwise compatible.
- By the Pairwise Compatibility Theorem (remember that?) they must then be jointly compatible
- So there must be a tree that contains them all.

The MR tree with 14-species primate mtDNA data



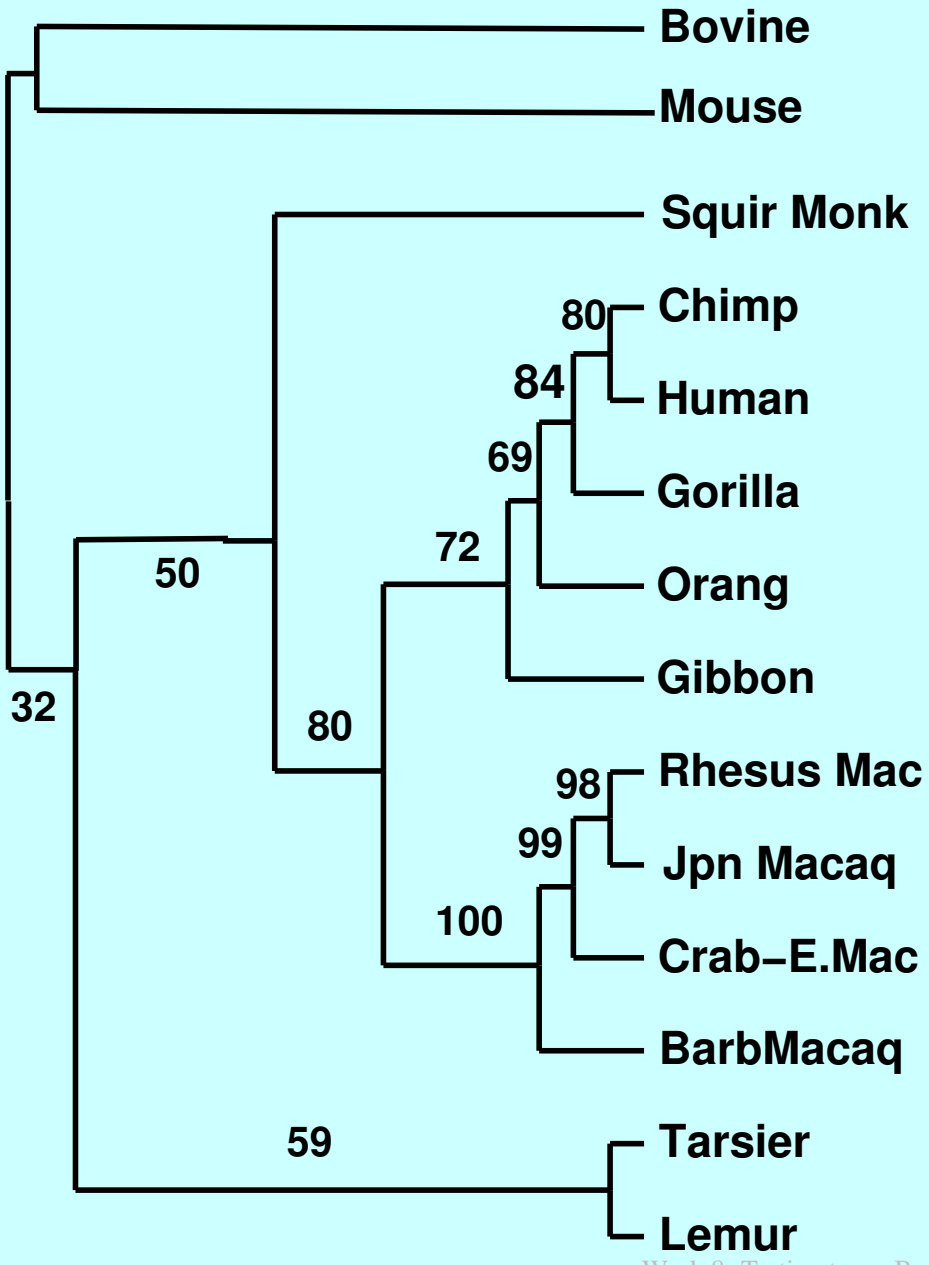
Potential problems with the bootstrap

1. Sites may not evolve independently
2. Sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions)
3. If do not know which branch is of interest at the outset, a “multiple-tests" problem means P values are overstated
4. P values are biased (too conservative)
5. Bootstrapping does not correct biases in phylogeny methods

Other resampling methods

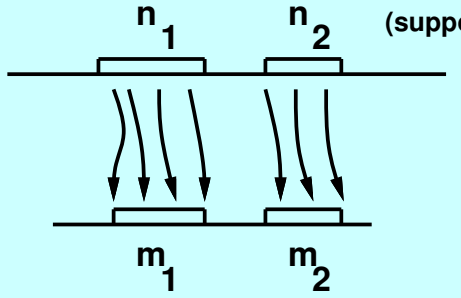
- Delete-half jackknife. Sample a random 50% of the sites, *without* replacement.
- Delete-$1/e$ jackknife (Farris et. al. 1996) (too little deletion from a statistical viewpoint).
- Reweighting characters by choosing weights from an exponential distribution.
- In fact, reweighting them by any exchangeable weights having coefficient of variation of 1
- Parametric bootstrap – simulate data sets of this size assuming the estimate of the tree is the truth
- (to correct for correlation among adjacent sites) (Künsch, 1989) Block-bootstrapping – sample n/b blocks of b adjacent sites.

With the delete-half jackknife



Bootstrap versus jackknife in a simple case

Exact computation of the effects of deletion fraction for the jackknife



(suppose 1 and 2 are conflicting groups)

n characters

n(1–δ) characters

We can compute for various n's the probabilities of getting more evidence for group 1 than for group 2

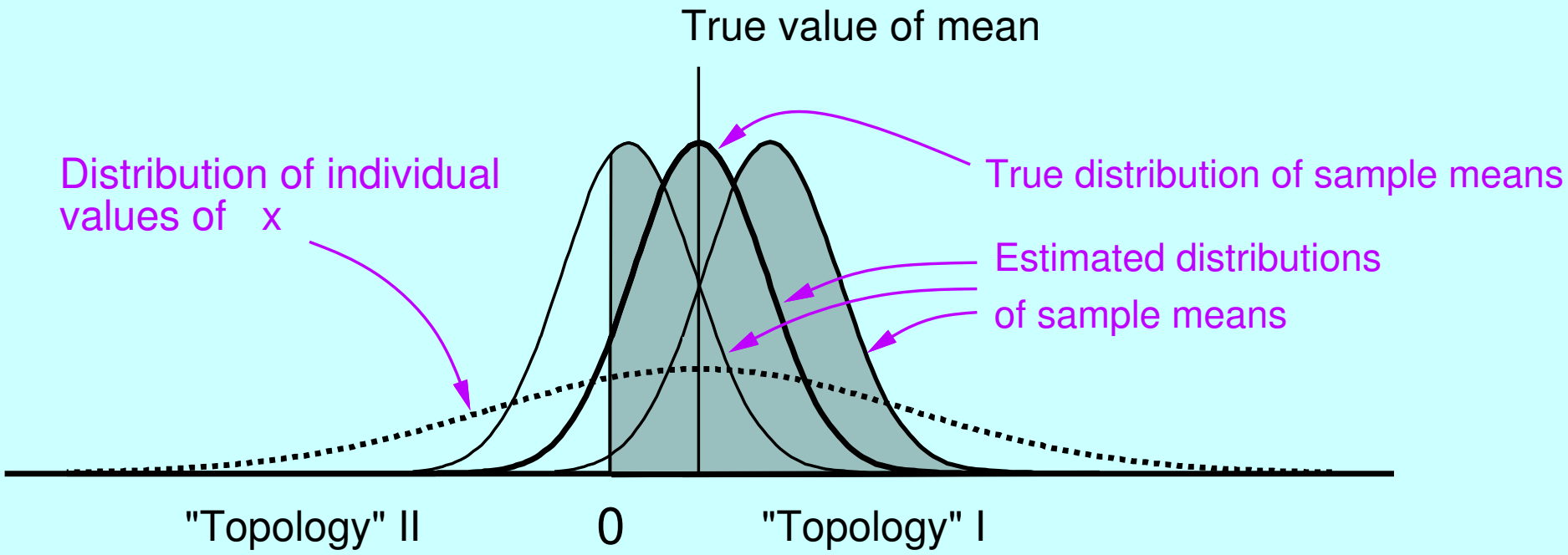
A typical result is for $n_1 = 10$, $n_2 = 8$, $n = 100$:

	Bootstrap	Jackknife $\delta = 1/2$	Jackknife $\delta = 1/e$
$\text{Prob}(m_1 > m_2)$	0.6384	0.5923	0.6441
$\text{Prob}(m_1 \geq m_2)$	0.7230	0.7587	0.8040
$\text{Prob}(m_1 > m_2) + \frac{1}{2}\text{Prob}(m_1 = m_2)$	0.6807	0.6755	0.7240

Probability of a character being omitted from a bootstrap

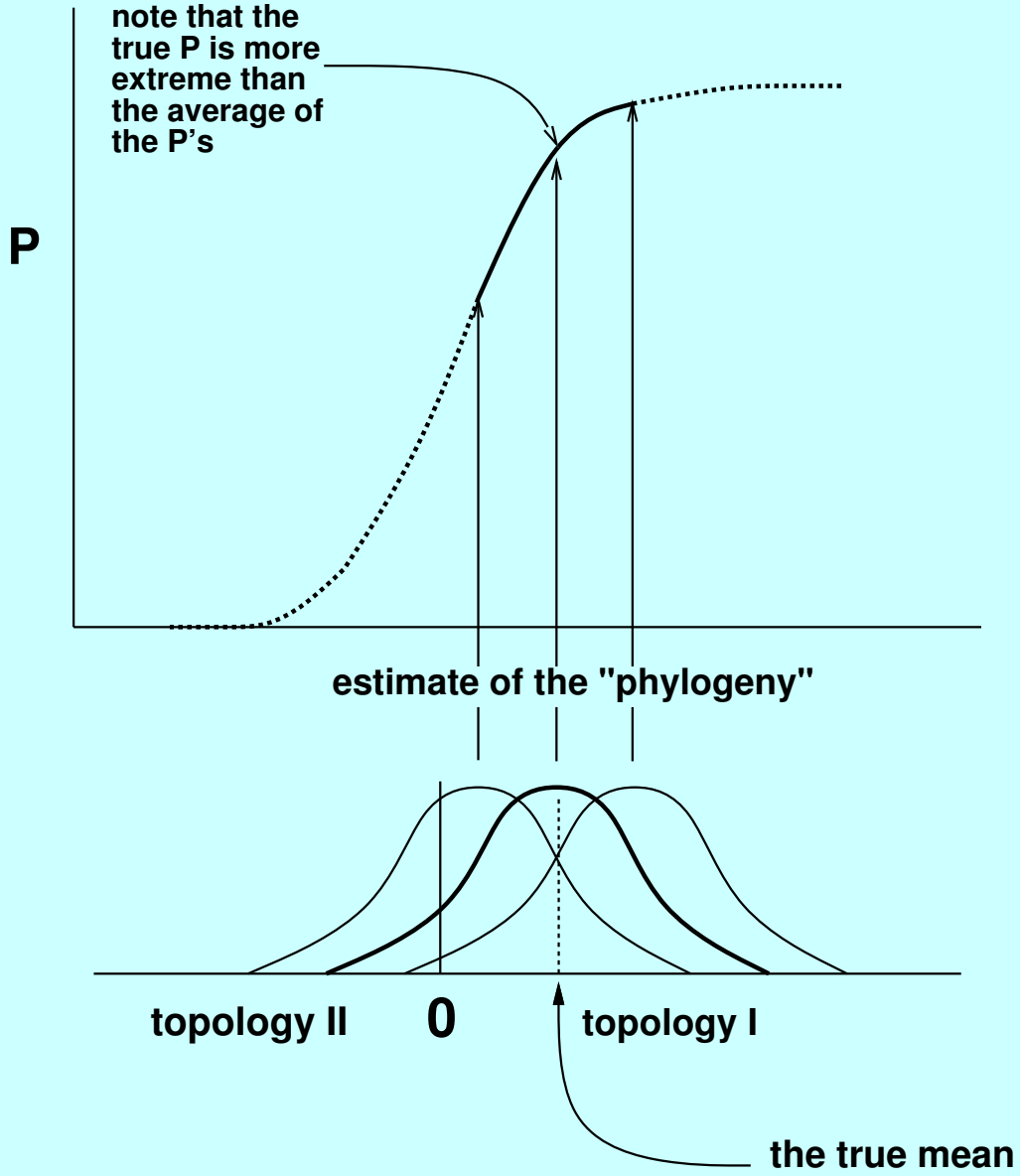
N	$(1-1/N)^N$						
1	0	11	0.35049	25	0.36040	100	0.36603
2	0.25	12	0.35200	30	0.36166	150	0.36665
3	0.29630	13	0.35326	35	0.36256	200	0.36696
4	0.31641	14	0.35434	40	0.36323	250	0.36714
5	0.32768	15	0.35526	45	0.36375	300	0.36727
6	0.33490	16	0.35607	50	0.36417	500	0.36751
7	0.33992	17	0.35679	60	0.36479	1000	0.36770
8	0.34361	18	0.35742	70	0.36524	∞	0.36788
9	0.34644	19	0.35798	80	0.36557		
10	0.34868	20	0.35849	90	0.36583		

A toy example to examine bias of P values

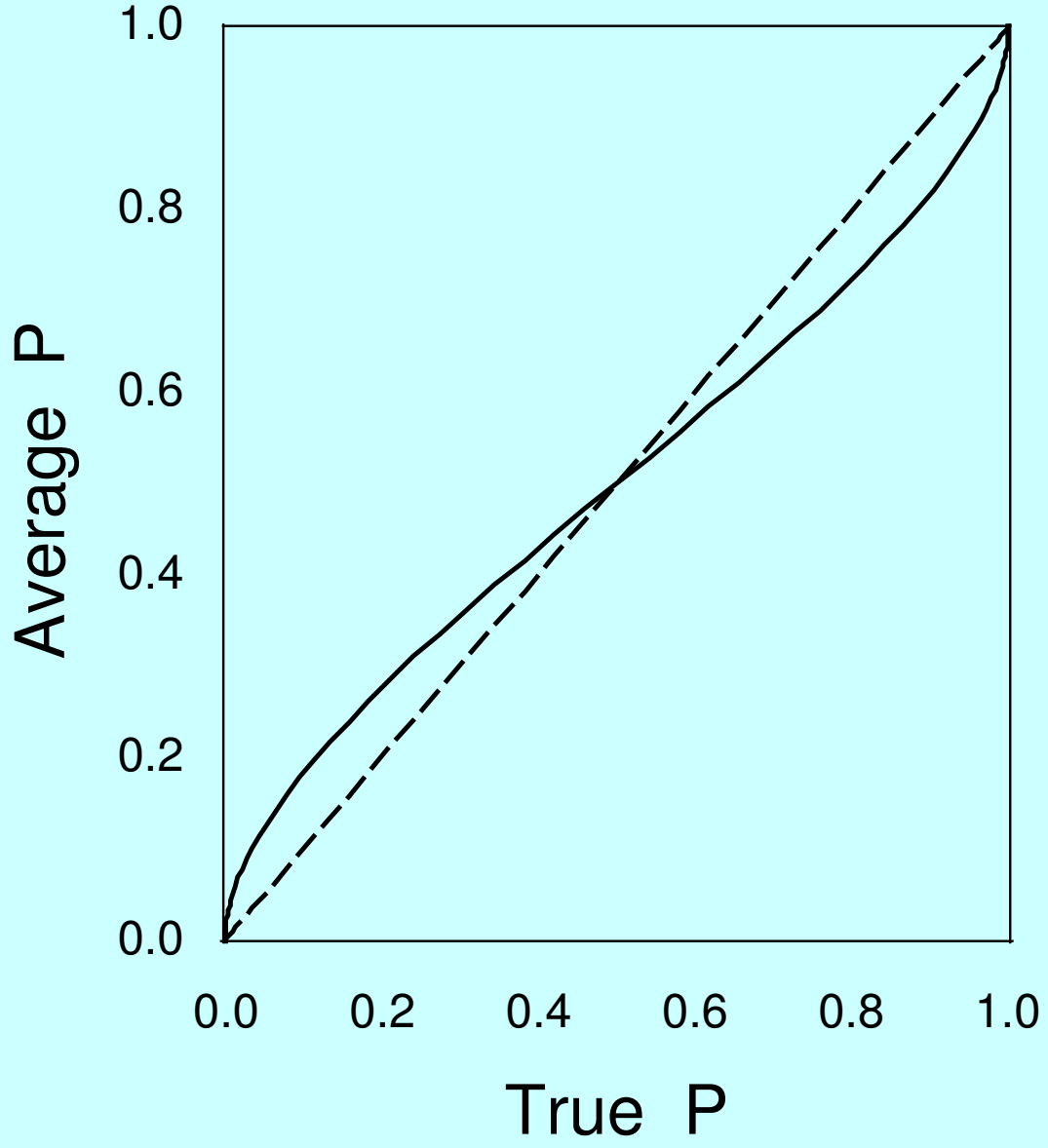


Assuming a normal distribution, trying to infer whether the mean is above 0, when the mean is unknown and the variance known to be 1

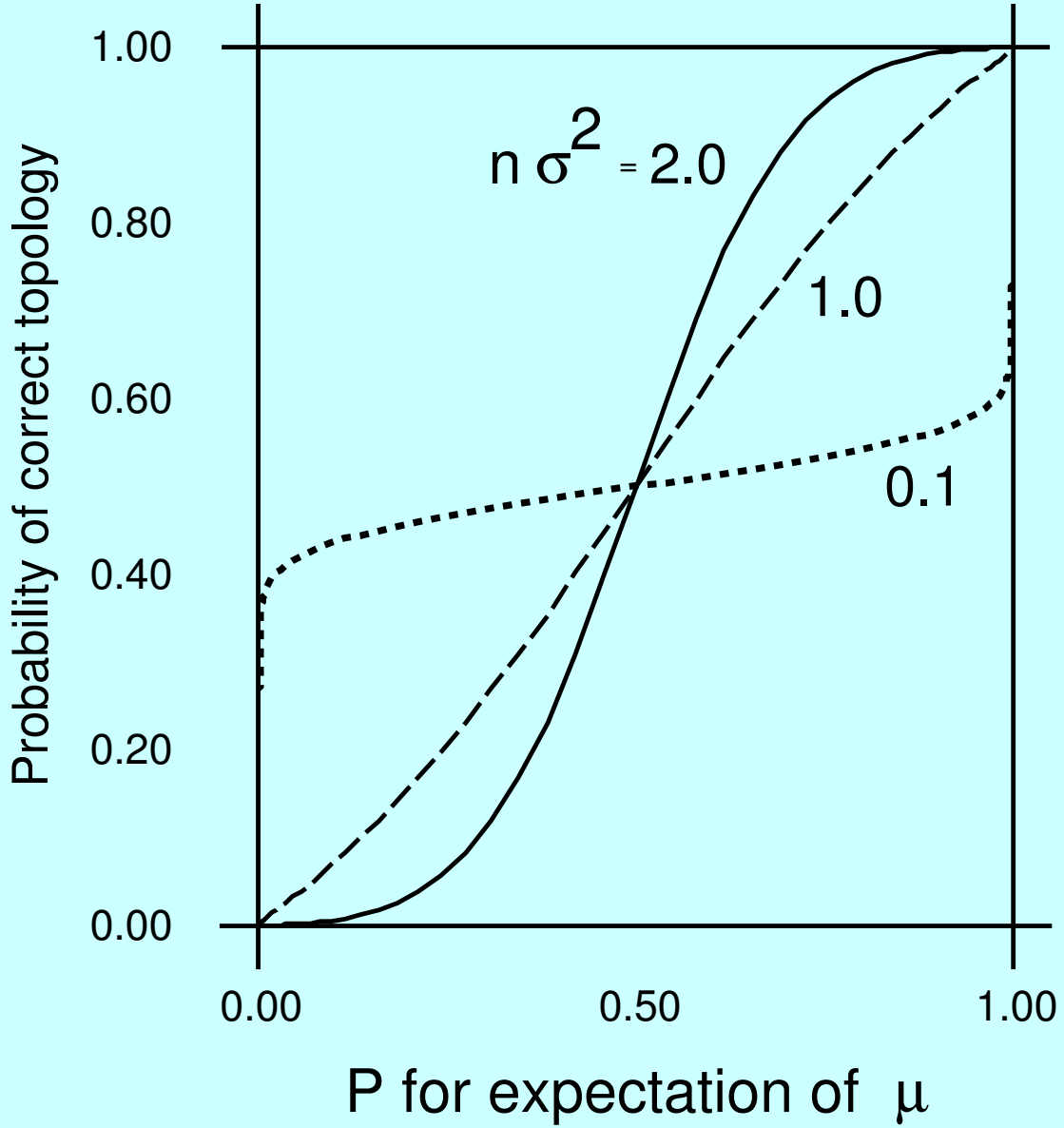
Bias in the P values



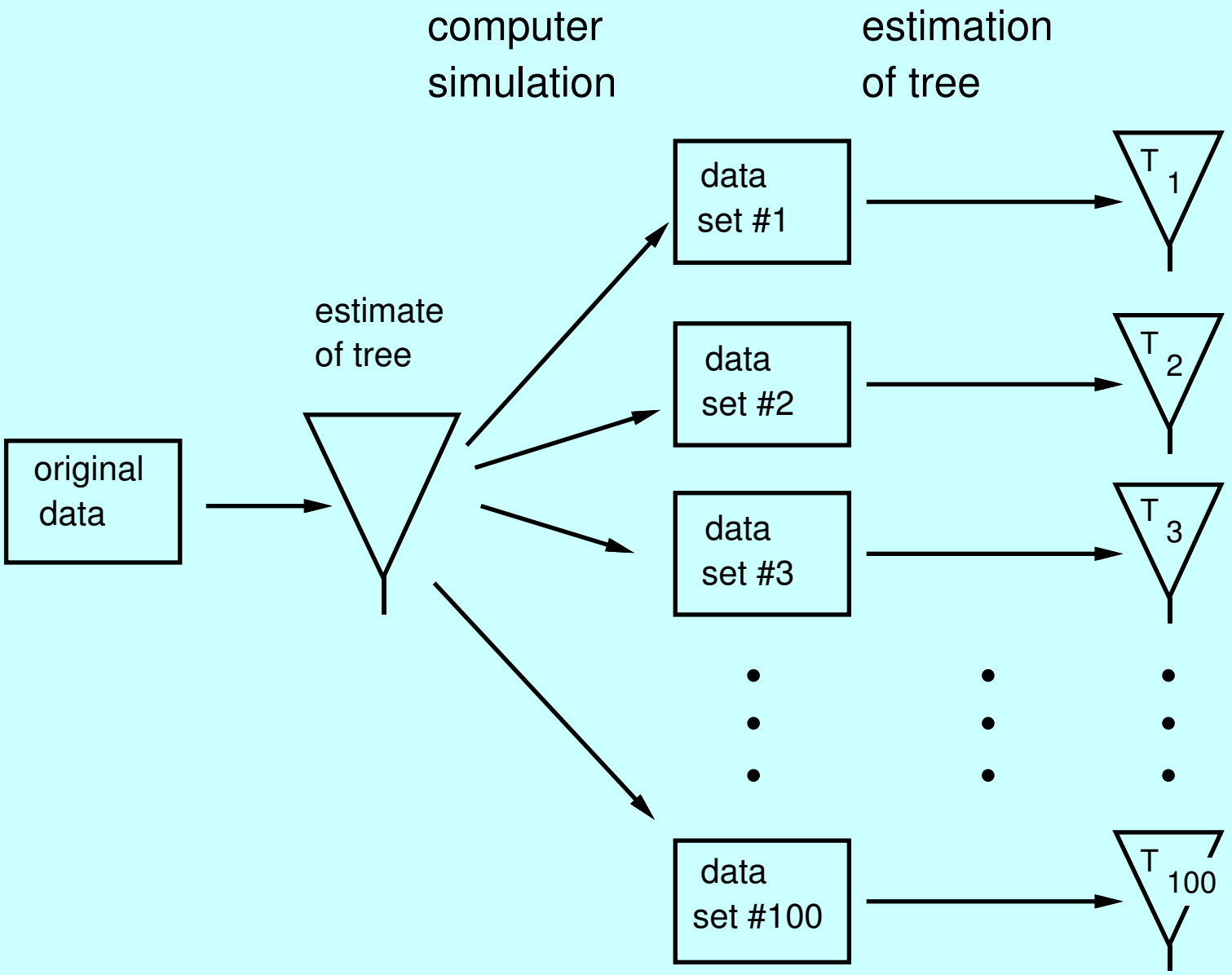
How much bias in the P values?



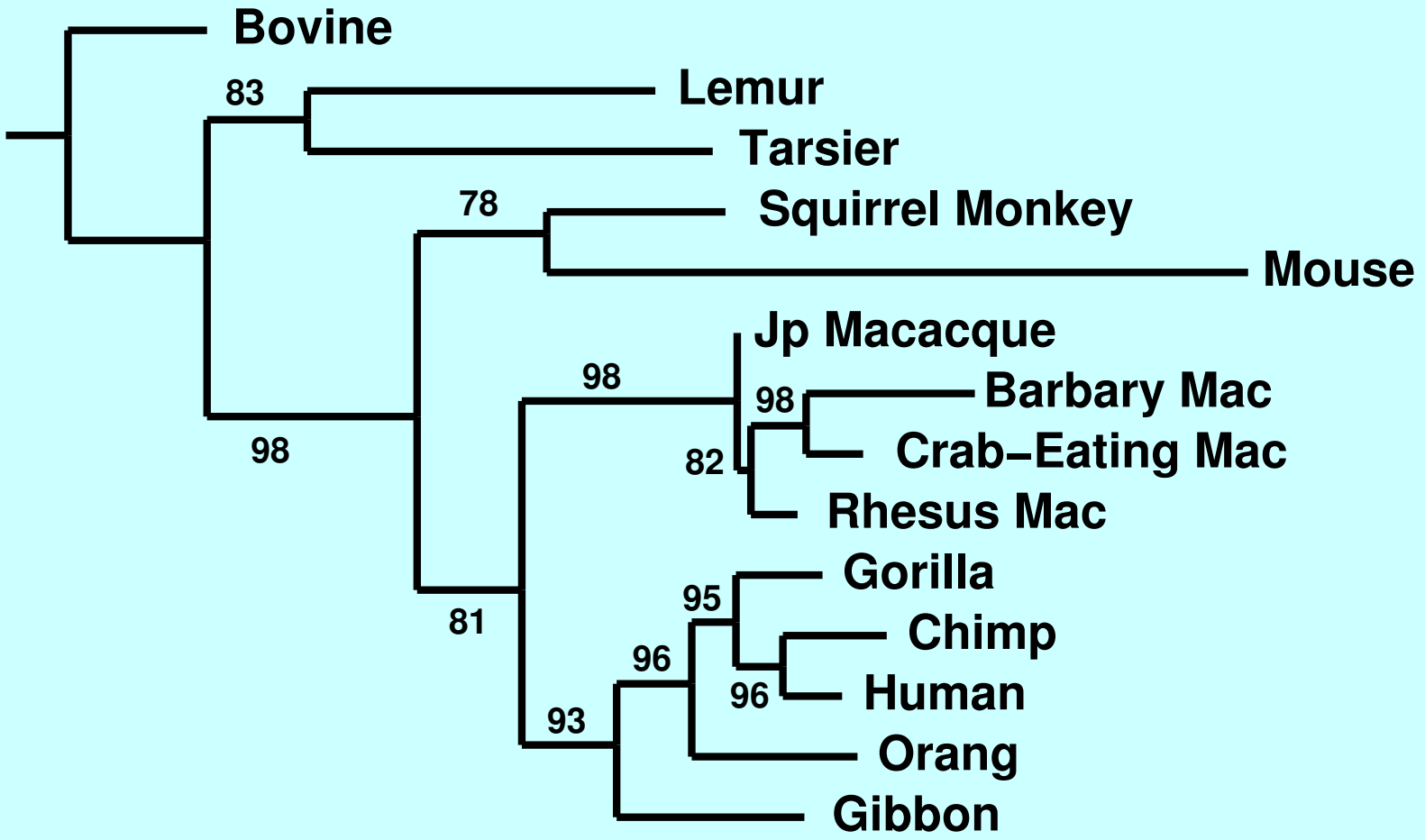
Bias in the P values with different priors



The parametric bootstrap

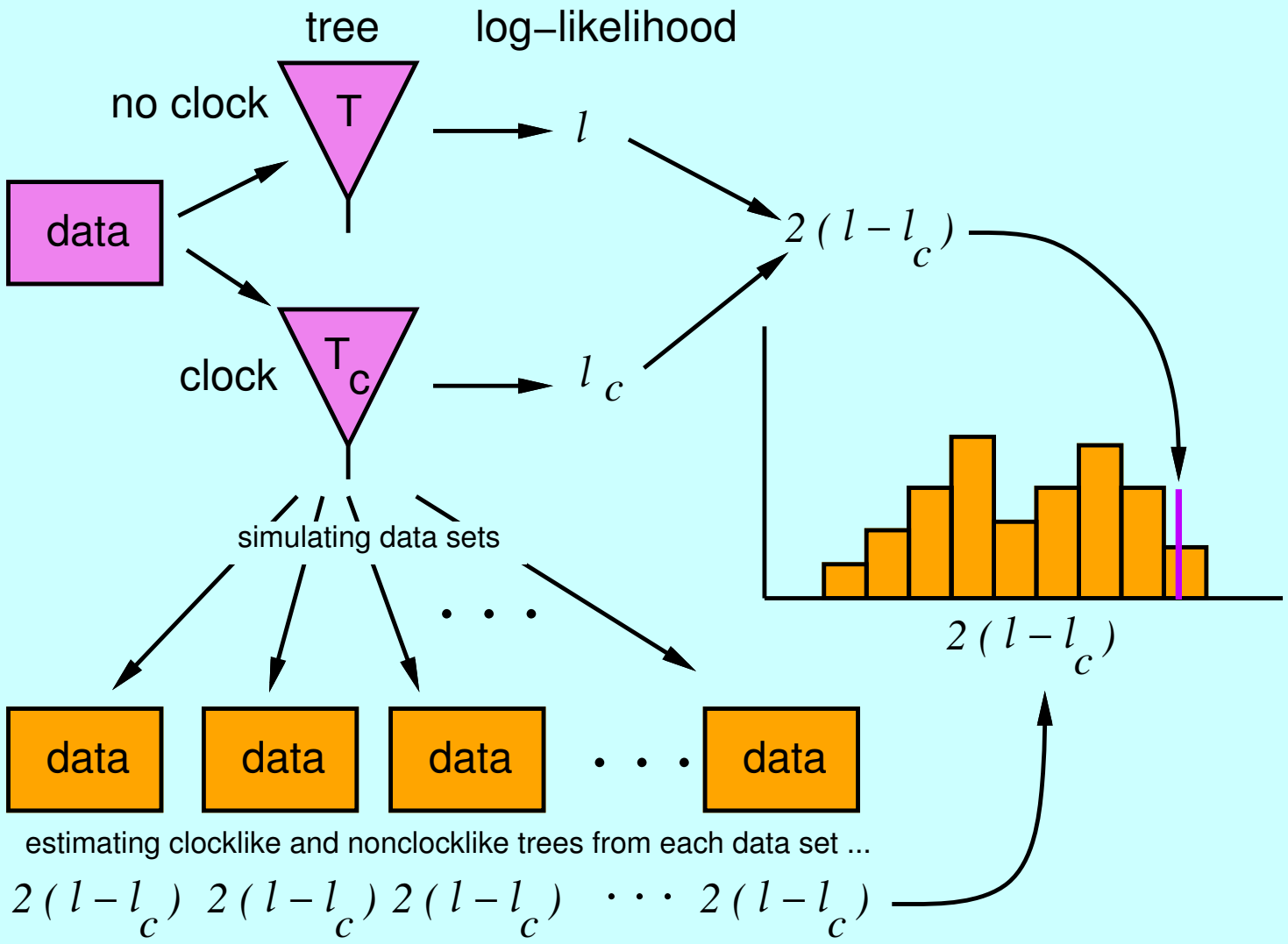


The parametric bootstrap with the primates data



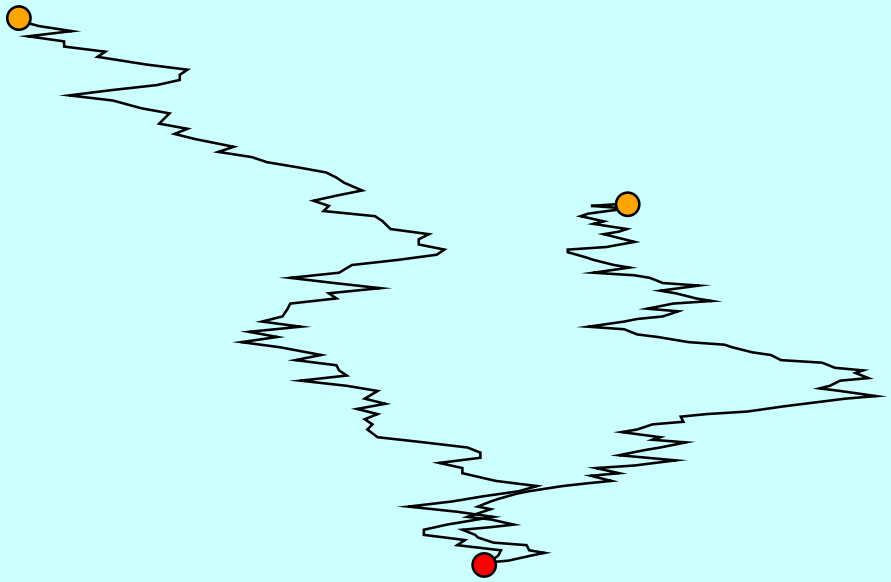
Goldman's test using simulation

(related to the "parametric bootstrap")

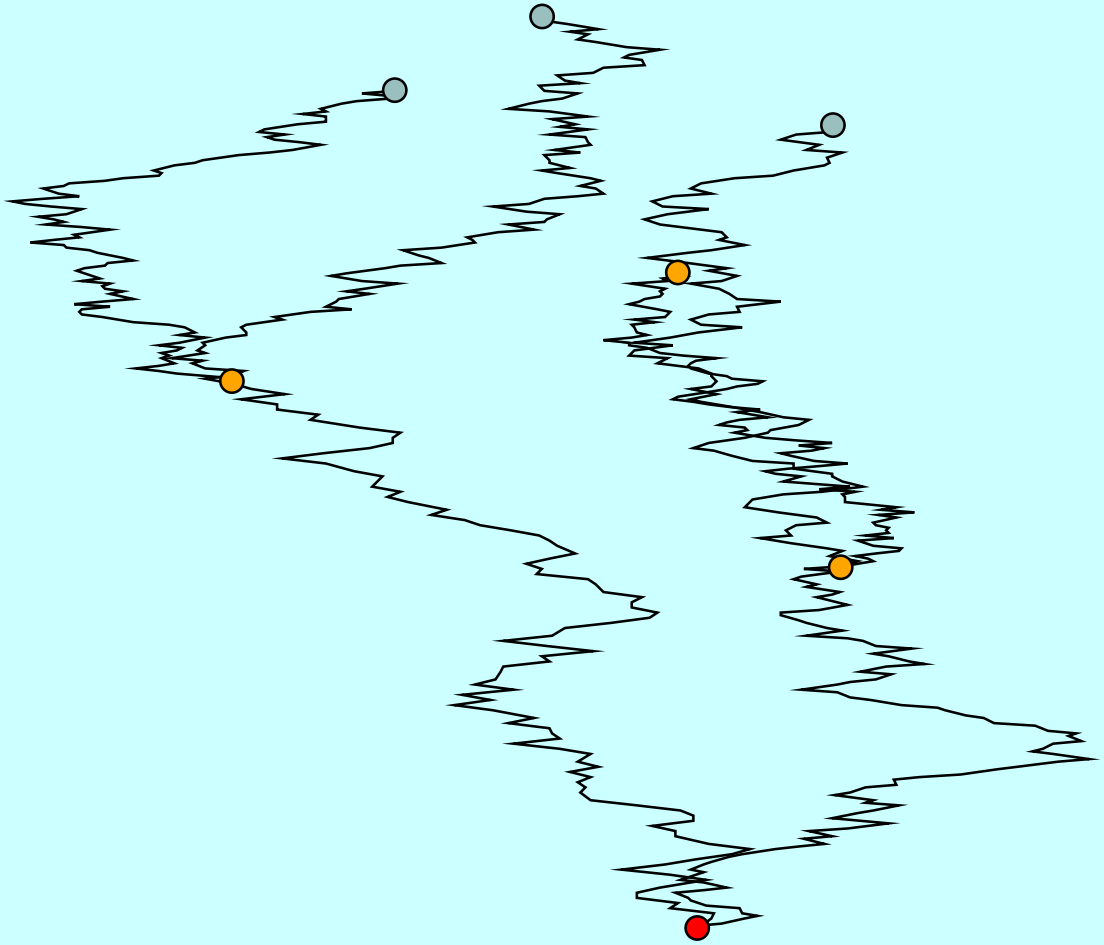


An outcome of Brownian motion on a 5-species tree

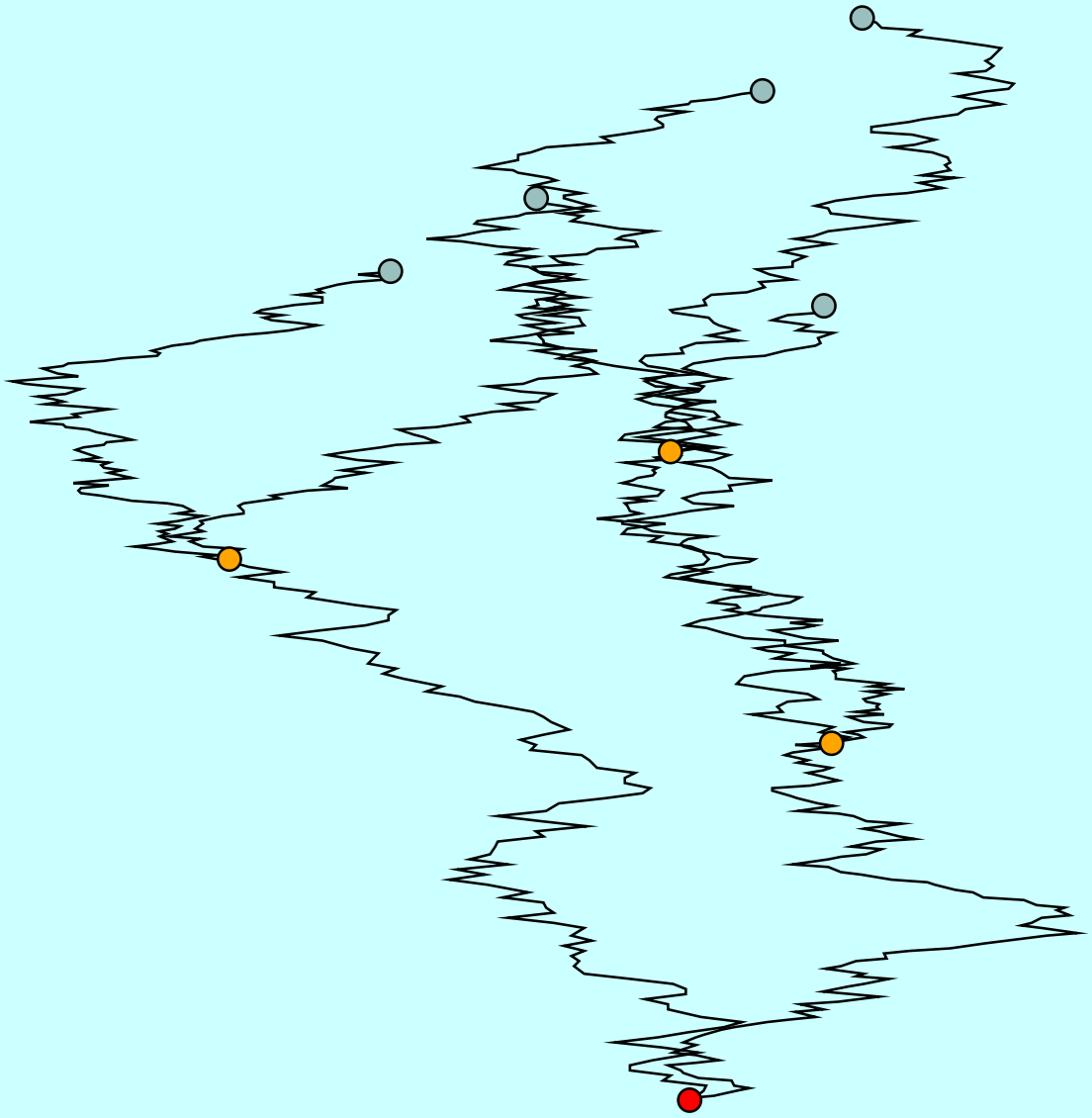
An outcome of Brownian motion on a 5-species tree



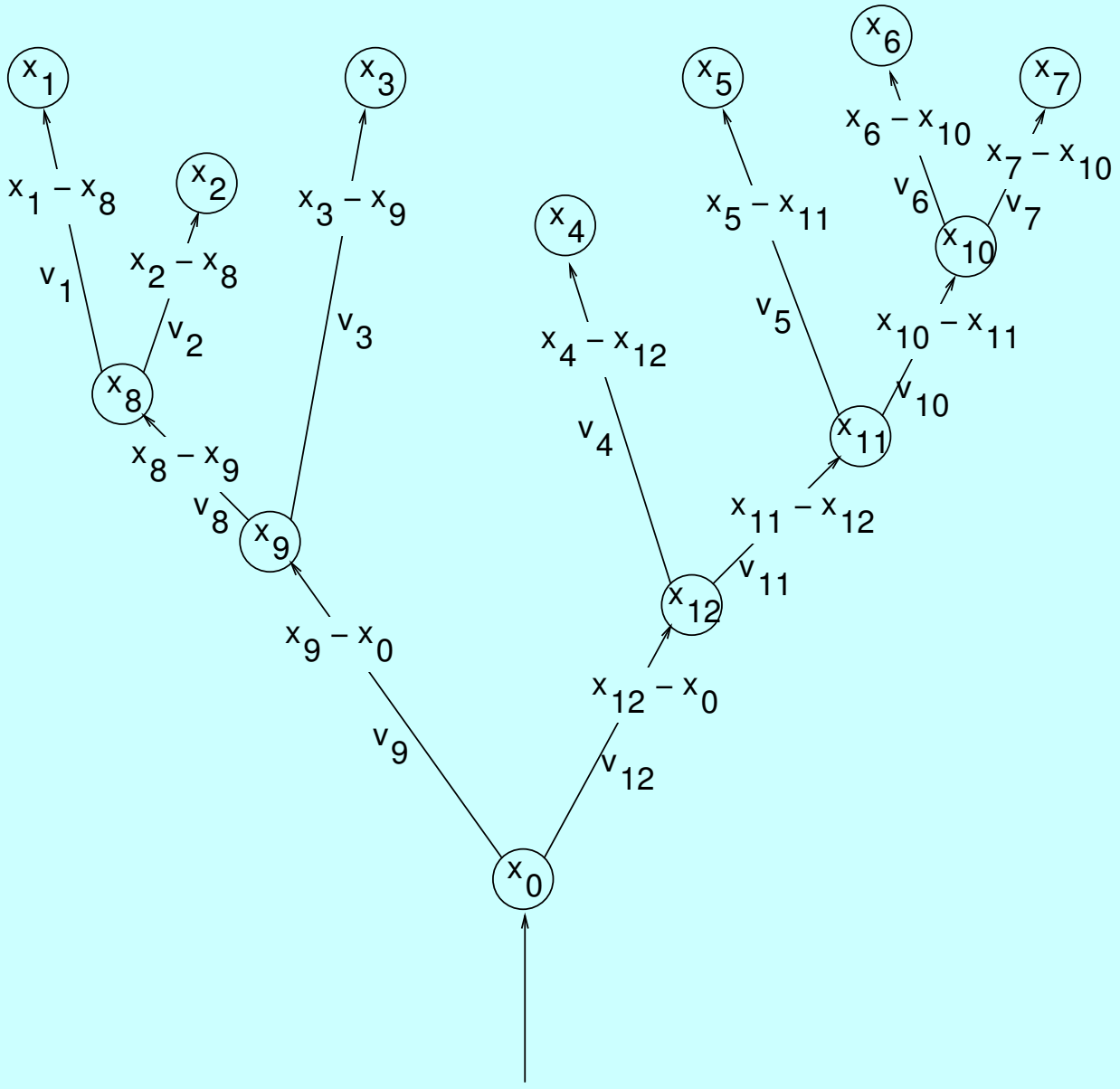
An outcome of Brownian motion on a 5-species tree



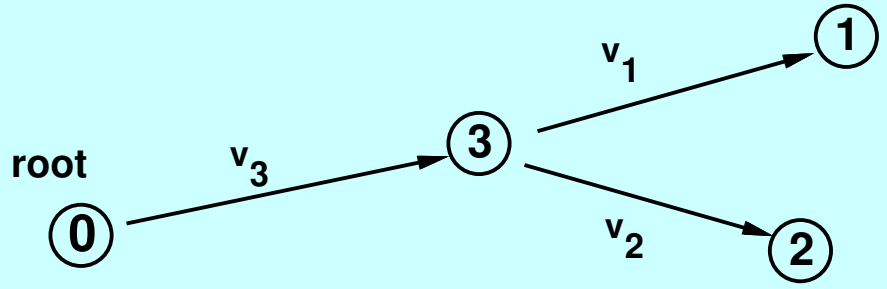
An outcome of Brownian motion on a 5-species tree



Brownian motion along a tree

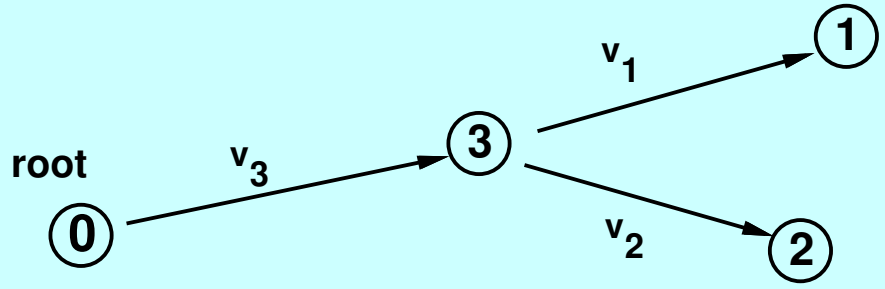


Distribution of tips on a tree under Brownian Motion



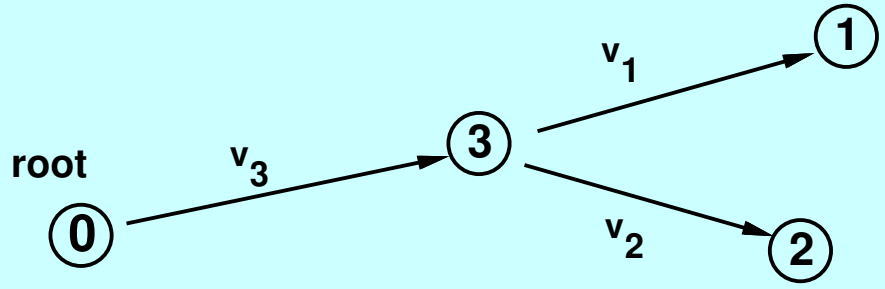
- Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances v_3 and v_1) so it is normally distributed with mean 0 and variance $v_3 + v_1$.

Distribution of tips on a tree under Brownian Motion



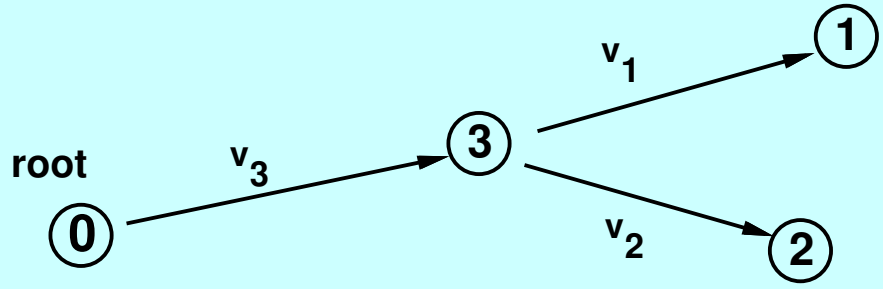
- Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances v_3 and v_1) so it is normally distributed with mean 0 and variance $v_3 + v_1$.
- Similarly for tip 2 (variance is $v_3 + v_2$).

Distribution of tips on a tree under Brownian Motion



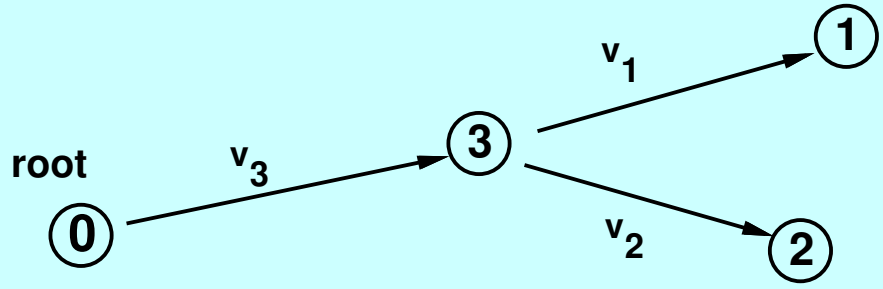
- Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances v_3 and v_1) so it is normally distributed with mean 0 and variance $v_3 + v_1$.
- Similarly for tip 2 (variance is $v_3 + v_2$).
- They share branch 3, and the change there affects both random variables. So they are not independent or uncorrelated.

Distribution of tips on a tree under Brownian Motion



- Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances v_3 and v_1) so it is normally distributed with mean 0 and variance $v_3 + v_1$.
- Similarly for tip 2 (variance is $v_3 + v_2$).
- They share branch 3, and the change there affects both random variables. So they are not independent or uncorrelated.
- Variance is the expectation of the square (of deviation from the mean), and covariance is the expectation of the product of those deviations, for the two variables.

Distribution of tips on a tree under Brownian Motion



- Tip 1 is the sum of two independent changes each of which is drawn from a normal distribution (with mean 0 and variances v_3 and v_1) so it is normally distributed with mean 0 and variance $v_3 + v_1$.
- Similarly for tip 2 (variance is $v_3 + v_2$).
- They share branch 3, and the change there affects both random variables. So they are not independent or uncorrelated.
- Variance is the expectation of the square (of deviation from the mean), and covariance is the expectation of the product of those deviations, for the two variables.
- In fact the covariance of the values at tip 1 and tip 2 is the variance of the shared term that is the same in both of them, so it is v_3.

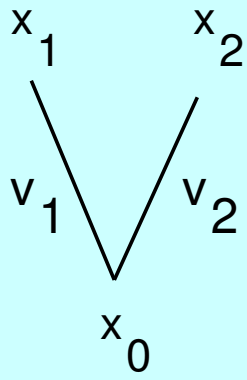
Covariances of species on the tree

$$\begin{bmatrix}
v_1 + v_8 + v_9 & v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_8 + v_9 & v_2 + v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 \\
v_9 & v_9 & v_3 + v_9 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & v_4 + v_{12} & v_{12} & v_{12} & v_{12} \\
0 & 0 & 0 & v_{12} & v_5 + v_{11} + v_{12} & v_{11} + v_{12} & v_{11} + v_{12} \\
0 & 0 & 0 & v_{12} & v_{11} + v_{12} & v_6 + v_{10} + v_{11} + v_{12} & v_{10} + v_{11} + v_{12} \\
0 & 0 & 0 & v_{12} & v_{11} + v_{12} & v_{10} + v_{11} + v_{12} & v_7 + v_{10} + v_{11} + v_{12}
\end{bmatrix}$$

Covariances are of form

$$\begin{bmatrix} a & b & c & 0 & 0 & 0 & 0 \\ b & d & c & 0 & 0 & 0 & 0 \\ c & c & e & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & f & g & g & g \\ 0 & 0 & 0 & g & h & i & i \\ 0 & 0 & 0 & g & i & j & k \\ 0 & 0 & 0 & g & i & k & l \end{bmatrix}$$

Likelihood under Brownian motion with two species



$$f\left(x;\mu,\sigma^2\right) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L \;=\; \prod_{i=1}^{p}\frac{1}{(2\pi)\sqrt{v_1 v_2}}\exp\left(-\frac{1}{2}\left[\frac{(x_{1i}\;-\;x_{0i})^2}{v_1}+\frac{(x_{2i}\;-\;x_{0i})^2}{v_2}\right]\right)$$

What happens if we estimate means and branch lengths?

Do we get the right answer if we estimate for each coordinate (each character) the value at the root and the branch lengths v_1 and v_2 ? Actually no.

Below, we will do this by finding values of these that maximize the likelihood, and show that the likelihood becomes infinite if either v_1 or v_2 approaches zero.

Even if we constrain there to be a clock, so $v_1 = v_2$ and look only at their sum $v_1 + v_2$ this turns out to be half as big as the truth, even with an infinite number of characters.

Why? The problem seems to be that we are estimating too many parameters. There is one parameter (the root value) for each character. So the ratio of data to parameters does not rise to infinity as we increase the number of parameters. In circumstances like this, likelihood methods can misbehave.

The solution: don't infer ancestors; use REML

We can eliminate these problems by:

1. Do not infer the states of the interior nodes.
2. Use only the relative positions of the tips. This eliminates the starting state at the root. It is REML, a variant of ML that loses almost no statistical power.

Minimizing for each character i

$$Q = \frac{(x_{1i} - x_{0i})^2}{v_1} + \frac{(x_{2i} - x_{0i})^2}{v_2}$$

so:

$$\frac{dQ}{dx_{0i}} = -2\frac{(x_{1i} - x_{0i})}{v_1} - 2\frac{(x_{2i} - x_{0i})}{v_2} = 0$$

and then:

$$\widehat{x}_{0i} = \frac{\frac{1}{v_1}x_{1i} + \frac{1}{v_2}x_{2i}}{\frac{1}{v_1} + \frac{1}{v_2}}$$

So that we have a maximum likelihood estimate of the starting value x_{0i} for each character.

The result is that

$$Q = \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2}$$

Likelihood after estimating initial coordinates

Substituting in our estimates of x_{0i}, we end up with

$$L = \frac{1}{(2\pi)^p (v_1 v_2)^{\frac{1}{2}p}} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{(x_{1i}-x_{2i})^2}{v_1+v_2}\right)$$

and this finally turns into:

$$\ln L = -p\ln(2\pi) - \frac{1}{2}p\ln(v_1 v_2) - \frac{1}{2}\sum_{i=1}^{p}\frac{(x_{1i}-x_{2i})^2}{v_1+v_2}$$

This actually goes to infinity as either v_1 or v_2 goes to zero! This is related to the problem that Edwards and Cavalli-Sforza had with their maximum likelihood method in 1964.

If there is a clock ...

If instead we constrain $v_1 = v_2$ because assume a clock:

$$\ln L = K' - p \ln(v_1 + v_2) - \frac{1}{2} \frac{D^2}{(v_1 + v_2)}$$

which leads to

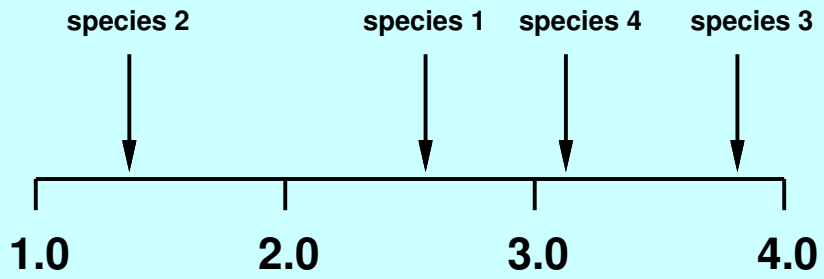
$$\widehat{v}_1 = \widehat{v}_2 = D^2/(4p)$$

(which is half as big as it should be!)

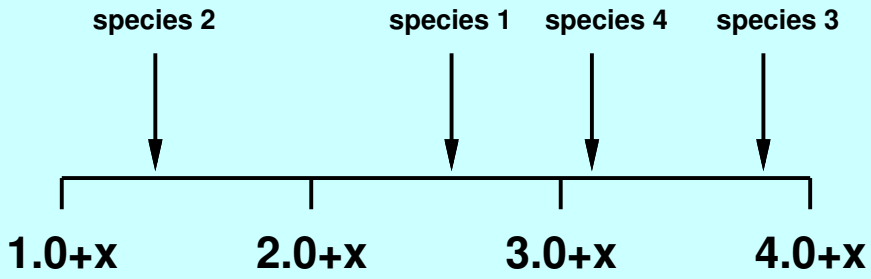
The number of parameters being estimated is $p + 1$, which rises as we consider more characters. The fact that the ratio of data to parameters does not rise without limit is the reason why likelihood misbehaves in this case.

The difference between ML and REML

Information we use for ML inference:



Information we use for REML inference:



Does it matter that we don't know x? It makes it unnecessary to estimate the starting value x_0, and that eliminates p parameters. It means that the ratio of data to parameters does then rise as we add characters.

Using only differences between populations (REML)

We assume that we have observed only the differences $x_{1i} - x_{2i}$, and not the actual locations on the phenotype scale. Then

$$L = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sqrt{v_1 + v_2}} \exp\left(-\frac{1}{2}\frac{(x_{1i} - x_{2i})^2}{v_1 + v_2}\right)$$

$$\ln L = K - \frac{p}{2}\ln(v_1 + v_2) + \frac{1}{2(v_1 + v_2)}\sum_{i=1}^{n}(x_{i1} - x_{i2})^2$$

Likelihood with two species using REML

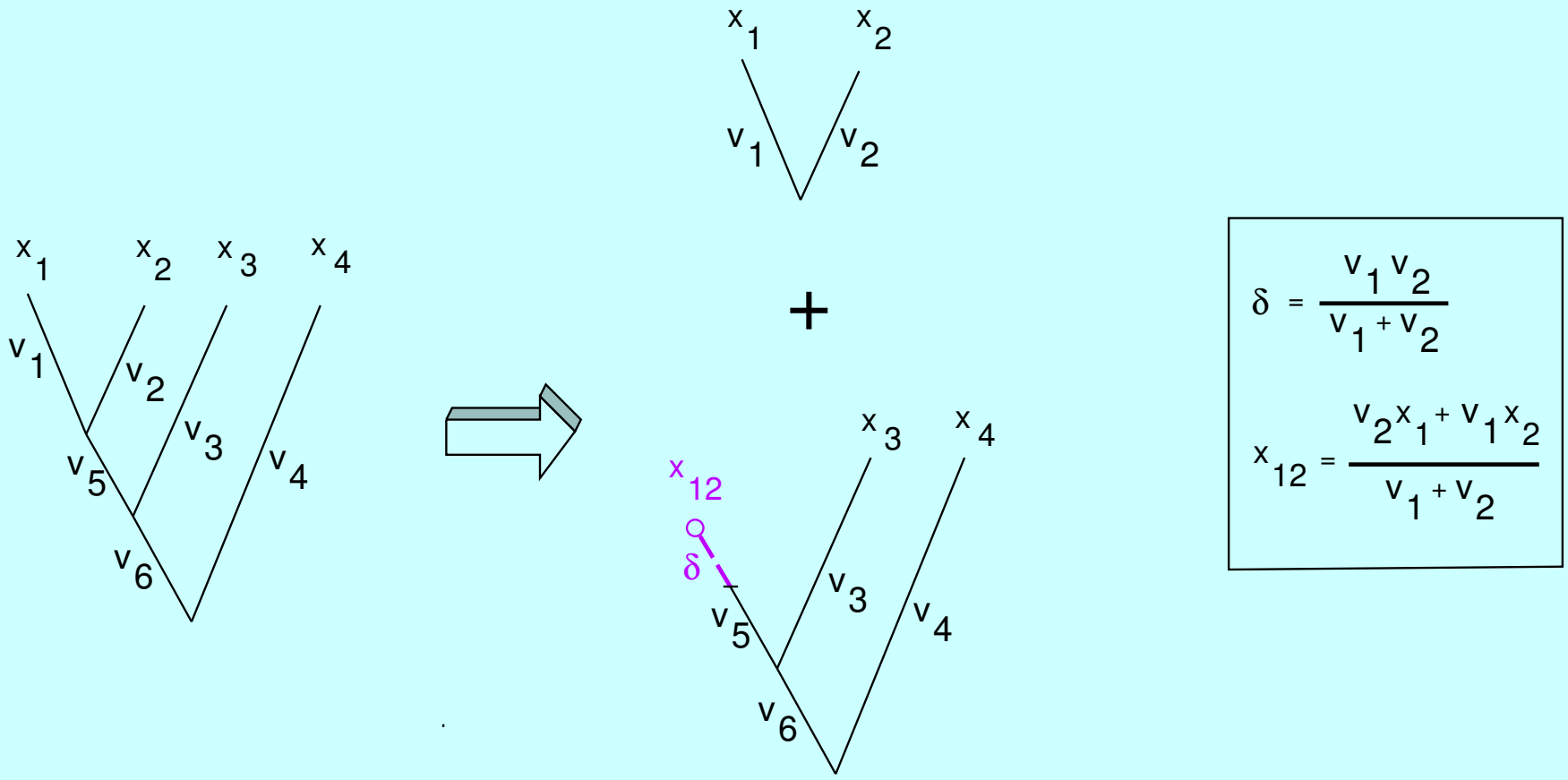
$$\ln L = K - \frac{p}{2} \ln (v_1 + v_2) + \frac{D^2}{2 (v_1 + v_2)}$$

$$\ln L = K - \frac{p}{2} \ln (v_T) + \frac{D^2}{2 v_T}$$

$$\widehat{v}_T = D^2/p$$

The number of parameters being estimated is 1 (it is the sum $v_1 + v_2$).
The number of parameters does not rise as we consider more characters.

“Pruning” a tree in the Brownian motion case



The likelihood for the tree is the product of the linkelihoods for these two trees. By repeatedly applying this we can decompose the tree into $n - 1$ independent two-species trees. Getting their likelihoods is easy.