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
Overview:

- Bootstrapping to estimate clade support
- Controversies about the interpretation of the bootstrap support
- The AU test
The bootstrap

- Estimate of $\theta$
- Empirical distribution of sample
- Bootstrap replicates
- Unknown true value of $\theta$
- Unknown true distribution

Distribution of estimates of parameters

Week 7: Bayesian inference, Testing trees, Bootstraps – p.33/54
The bootstrap for phylogenies

Original
Data
sequences
sites
Bootstrap
sample
#1
Bootstrap
sample
#2
Estimate of the tree,
#1
Bootstrap estimate of
the tree,
#1
Bootstrap estimate of
the tree,
#2
Bootstrap estimate of
the tree,
#2
Estimate of the tree,
(and so on)
sample same number
of sites, with replacement
sample same number
of sites, with replacement
Week 7: Bayesian inference, Testing trees, Bootstraps – p.34/54
Bootstrapping: first step

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td></td>
<td>G</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>weights</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

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

Sum of weights equals $k$ (i.e., each bootstrap dataset has same number of sites as the original)
Bootstrapping: second replicate

<table>
<thead>
<tr>
<th>weights</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>0</th>
<th>...</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>...</td>
<td>G</td>
</tr>
</tbody>
</table>

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

Note that weights are different this time, reflecting the random sampling with replacement used to generate the weights.
Bootstrapping: 20 replicates

<table>
<thead>
<tr>
<th>1234</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>-**-</td>
<td>75.0</td>
</tr>
<tr>
<td>-***</td>
<td>15.0</td>
</tr>
<tr>
<td>-***</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Note: usually at least 100 replicates are performed, and 500 is better.
Bootstrap support for branches can be displayed:

- on the best estimate of the tree (on the MP, OLS, or ML tree)
- on a majority-rule consensus tree (see Berry and Gascuel, 1996)
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
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).

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

![Graph showing the comparison between Bayesian and Bootstrap proportions](image)

from Wilcox et al. (2002)
What did Efron et al. (1996) say?

We can use a Bayesian model to show that $\hat{\alpha}$ is a reasonable assessment of the probability that $R_1$ contains $\mu$. Suppose we believe a priori that $\mu$ could lie anywhere in the plane with equal probability. Then having observed $\hat{\mu}$, the a posteriori distribution of $\mu$ given $\hat{\mu}$ is $N_2(\hat{\mu}, I)$ exactly the same as the bootstrap distribution of $\hat{\mu}^*$. In other words, $\hat{\alpha}$ is the a posteriori probability of the event $\mu \in R_1$, if we begin with an “uninformative” prior density for $\mu$. 

![Diagram](attachment:boundaryDiagram.png)
Kim (2000) view of tree space

Tree Model

Tree Estimation Method

Finite Data Sets
Pattern Frequency Space With Observed Data

\[ P(1100) \]

\[ \hat{\mu} \]

\[ P(1010) \]

\[ P(1001) \]

\[ P(1010) \]

\[ P(1001) \]
Bootstrapping in Pattern Frequency Space

P(1100)

P(1010) P(1001)

P(1010) P(1001)
Posterior Densities in Pattern Frequency Space
Posterior Densities in Pattern Frequency Space
What did Efron et al. (1996) say (and mean)?

- the “uninformative” prior density is a uniform prior over all of pattern 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 $\infty$),

- 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.
Bootstrap proportions have been characterized as providing:

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

\[ N = 100 \text{ and } H = 60 \]

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 number of heads.

Frequency
0 20 40 60 80 100
0 1000 2000 3000 4000

\[ P\text{-value} \approx 0.029 \]
If we bootstrap we get a sense of the variability in our estimate, but we can also get a tail probability for $\Pr(p^{(\text{boot})} \leq 0.5)$.
\[ \Pr(p^{(\text{boot})} \leq 0.5) \approx 0.027 \]
A simulation of the null distribution of the # heads

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

$\mu^\dagger$ is least-favorable condition (LFC) point in $R_0$

Case 1 P-value $<$ the P-value in Case 2
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

\[ \hat{\mu} \quad \mu^\dagger \]

\[ R_0 \quad R_1 \]

Case 4

\[ \hat{\mu} \quad \mu^\dagger \]

\[ R_0 \quad R_1 \]

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

\( \mu^\dagger \) is least-favorable condition (LFC) point in \( R_1 \)
\[\hat{\mu}\] is the best point calculated from the data
\[\mu_\dagger\] is least-favorable condition (LFC) point in \(R_0\)
green areas are the tails - they correspond to values of the test statistic more extreme than \(\hat{\mu}\) (relative to that \(\mu \in R_1\))
Case 3 P-value > the P-value in 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 $c$:

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

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

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

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


Shimodaira, H. (2002). An approximately unbiased test of

