Chapter 20

Bootstrap, jackknife, and permutation tests

As likelihood requires us to believe the probability model of evolution, it may underestimate the amount of uncertainty about the tree. It would be desirable to have a less parametric approach to testing phylogenies. Bootstrap, jackknife, and randomization tests are one way to be less dependent on a complete parametric model. They use empirical information about the variation from character to character in evolutionary processes. A second reason for using these resampling techniques is that they allow us to infer the variability of parameters in models that are too complex for easy calculation of their variances.

Bootstrap and jackknife tests on phylogenies started with the work of Mueller and Ayala (1982), who used a jackknife approach to estimating the variance of the length of a branch in a UPGMA phylogeny from gene frequency data. This was followed by my own paper on the bootstrap (1985b) and those of Penny and Hendy (1985, 1986), who used random partitioning of the characters into two halves.

The bootstrap and the jackknife

The jackknife and bootstrap are statistical techniques for empirically estimating the variability of an estimate. They differ, but are of the same family of techniques. The jackknife, which is the older of the two, involves dropping one observation at a time from one’s sample, and calculating the estimate each time. The variability of the estimate is then inferred from the rather small variations that this causes, by an extrapolation. The bootstrap involves resampling from one’s sample with replacement, and making a fictional sample of the same size. We start by giving a general explanation of the bootstrap, and then consider how it can be applied to phylogenies.
The bootstrap was invented by Bradley Efron (1979) as a general-purpose statistical tool analogous to the jackknife. Figure 20.1 shows a diagram of the method. Imagine that we have some data points $x_1, x_2, x_3, \ldots, x_n$ that are drawn independently from a distribution $F(\theta)$, that depends on an unknown parameter, $\theta$. From them we are computing an estimate $t(x_1, x_2, \ldots, x_n)$ of the parameter $\theta$. We would like to know the variability of the distribution of these estimates. If we knew the
family of distributions from which $F$ came, and if the estimator $t(x)$ were mathematically tractable, then we could know the distribution of estimates and how it depended on the true $\theta$. For instance, when $F(\theta)$ is a normal distribution with mean $\theta$ and variance 1, and $t(x)$ is simply the sample mean, we know precisely what the distribution of estimates of $\theta$ is for every possible value of $\theta$. (It is normal, with mean $\theta$, and variance $1/n$.) That helps greatly in understanding what an estimate of $\theta$ implies.

However, we may not know the distribution $F$, or the estimator $t(x)$ may not be mathematically tractable. Efron's insight was that in this case, if the sample size $n$ is sufficiently large, we can consider the empirical distribution of data in our sample (which we can call $\hat{F}$) to estimate the true distribution $F$. Of course, the overall estimate of $\theta$ is not precisely correct, but the kinds of variation that the collection of values $x_1, x_2, \ldots, x_n$ display should be typical of the variation we would see in any large sample from the true distribution.

We would like to know what variation we would see in the estimate, $\hat{\theta}$, if we drew new data sets of size $n$ from the unknown distribution and analyzed them in the same way. The bootstrap infers this variation by using our current data set, by drawing new data sets not from $F$ but from the empirical distribution $\hat{F}$ in our data. Drawing a sample of size $n$ from the empirical distribution is the same as drawing a sample of points $x_1^n, x_2^n, \ldots, x_n^n$ from the existing data, drawing them independently, and sampling with replacement. If we instead sampled $n$ points without replacement, we would simply end up drawing each point once, and we would back get our original data, although the points would be in a different order. This would not result in a different estimate of $\theta$. But drawing with replacement means that points in the original data may be sampled different numbers of times. Some may be sampled twice, some once, some not at all (and some larger numbers of times). The numbers of times each one is drawn, $m_1, m_2, \ldots, m_n$, is a sample from a multinomial distribution with $n$ classes that have equal probabilities of being drawn.

This sample $x^*$ is called a bootstrap replicate. Each such replicate can be analyzed using the estimator $t$ to get $\hat{\theta}^* = t(x^*)$. To get a picture of the variation of the estimates $\theta$, we draw many different bootstrap replicates and infer $\theta$ from each one. The amount and kinds of variation in the resulting cloud of estimates of $\theta$ is then taken to be typical of the kinds of variation we would see if we could somehow sample new data sets from the unknown distribution $F$. For many well-behaved distributions and many well-behaved estimators $t(x)$ there are theorems assuring us that this picture of the variability of $\hat{\theta}$ will be accurate, if $n$ is large and if a large number of bootstrap replicates are taken.

**Bootstrapping and phylogenies**

To use the bootstrap to assess the uncertainty of our estimate of the phylogeny, the data should be a series of independently sampled points. We typically have,
instead, a matrix of species × characters. We cannot consider the species to be independent samples — they come instead as tips on an unknown phylogeny, some closely related to each other. In fact, the whole point of our analysis is to discover this structure. The characters (or sites) are a better candidate for being independent samples. If different characters evolve independently on the same phylogeny, they will satisfy the independence assumptions of the bootstrap, since the outcome of evolution at each character cannot be predicted from that in neighboring characters. Of course, evolutionary outcomes and processes in different characters may be related, in which case the independence assumption is incorrect. We return this subject later in this chapter.
To apply the bootstrap, we sample whole characters from the set of \( n \) characters, with replacement, and do so \( n \) times. The result is a data matrix with the same number of species and the same number of characters as in our original data matrix. Some of the original characters may have been sampled several times, others left out entirely. Figure 20.2 shows the process. For each data matrix we use our favorite phylogeny method to infer the phylogeny. The method may be a parsimony, distance matrix, or likelihood method. If a distance matrix method is used, the resampling occurs on the original character data (or sequences) before the distance matrix is computed. We end up with a collection of different estimates of the phylogeny. Some methods may give us more than one estimate of the phylogeny (parsimony methods, for example, often find multiple trees that are tied for best). In such cases we can consider that if 10 tied estimates are found for one bootstrap replicate, we consider each to be one-tenth of a tree, so that the results from that bootstrap replicate are not overemphasized when the trees are combined.

**The delete-half jackknife**

Other resampling methods are possible, and may have approximately equivalent behavior. The delete-half jackknife (e.g., Wu, 1986; Felsenstein, 1985b) is one, which has many of the same properties as the bootstrap. It involves sampling, not \( n \) times with replacement, but \( n/2 \) times without replacement. Thus we are taking a random half of the characters. Actually, if there are \( r \) parameters being estimated for each sample, we are supposed to take a random fraction \((n + r - 1)/2\) of the characters. For large \( n \) this will not make much difference, and it is hard to know what the value of \( r \) is for a phylogeny. The matter needs a closer examination.

One way to put the bootstrap and the delete-half jackknife into a common context is to consider them as randomly reweighting the data. Drawing a bootstrap sample is equivalent to putting new weights on the original data, with the weight on character \( i \) being the number of times, \( m_i \), that it is sampled in the bootstrap. As noted above, the weights \( m_i \) have a multinomial distribution, with \( n \) trials and equal probabilities for all \( n \) characters. It is not hard to show that the mean weight of a character is then 1, and the variance of the weight is \( 1 - 1/n \). Their coefficient of variation (the ratio of the standard deviation to the mean) is then \( \sqrt{1 - 1/n} \), which is nearly 1.

A jackknife that deletes a fraction \( f \) of the characters can be thought of as weighting the deleted characters 0 and the included characters 1. This implies a mean weight per character of \( 1 - f \) and a variance of \( f(1 - f) \). The coefficient of variation is then \( \sqrt{f/(1 - f)} \). When \( f = 1/2 \) we have a coefficient of variation of 1. It can be shown that any random weighting scheme that achieves the same coefficient of variation will also approximate the bootstrap.

It is not clear whether the delete-half jackknife has any substantial advantages over the bootstrap.

Farris et al. (1996) have advocated using a delete-\( 1/e \) jackknife together with a parsimony estimate of the phylogeny (their "Parsimony Jackknife"). \( 1/e \) is
0.36788, so this amounts to deleting substantially fewer characters, so that groups will appear to have more support than they would under a delete-half jackknife (or a bootstrap). We can evaluate this method by checking its behavior in a case where exact computations can be done. Suppose that we have 100 characters, 10 of which back group I, and 8 of which back group II, these two groups being incompatible. The other 82 characters do not discriminate between the two alternatives. We can calculate by exact enumeration of outcomes, calculating the probability of each, that a bootstrap sample will favor the first group 0.63836 of the time, with a tie 0.08461 of the time. It seems fairest to count the resampling as favoring the first group half of the time when there is a tie. This will be 0.6386 + (0.08461)/2 = 0.68066 of the time. If we do a delete-half jackknife, the corresponding number is 0.67555, while in a delete-1/e jackknife that samples 63 characters it is 0.72402. Thus the delete-half jackknife gets results much more consistent with the bootstrap.

Farris et al. chose delete-1/e based on the behavior when all the support is for group I. If two characters support group I and none group II, then the probability of favoring group I is 0.86738 for the bootstrap, 0.75253 for the delete-half jackknife, and 0.86545 for the delete-1/e jackknife. However, this match between the delete-1/e jackknife and the delete-1/e jackknife vanishes quickly as more characters favor group II. With just a few of them, the delete-half jackknife becomes closer to the bootstrap. Of course, if the bootstrap is to be the standard, this speaks in favor of using it instead.

The bootstrap and jackknife for phylogenies

Once we use the bootstrap (or the jackknife) to resample characters, we will have a cloud of trees, the results of estimating the phylogeny for each bootstrap or jackknife replicate. In the simple case of estimating a real-valued parameter, we can make a histogram of the estimates. How are we to do this with phylogenies? They have discrete topologies, but continuous branch lengths. We could use the bootstrap to make a histogram of branch lengths, but only if the branch in question existed in all of our estimates of the phylogeny. We might then, for example, make an interval estimate of the branch length by finding the upper 95% of the branch length histogram, so that we could infer a lower limit on the branch length.

If this lower limit were positive, we would then be asserting the existence of that branch. But suppose that the branch is missing in some of the bootstrap (or jackknife) estimates of the phylogenies. It seems reasonable to assume that those cases can be lumped with ones that have a zero branch length for this branch. If we do that, then we can assign the probability \( P \) to the branch if a fraction \( P \) of the bootstrap (or jackknife) replicates have the branch present. In cases where there are several tied trees in a bootstrap (or jackknife) estimate, some with the branch and some without, we can count each one as conferring fractional support for the branch. An alternative, and equivalent, way of looking at this is to imagine an indicator variable that is 1 if the branch exists in the bootstrap (or jackknife) esti-
Figure 20.3: A set of five trees and their majority-rule consensus tree, with the percentage of support for each interior branch shown. Note that the majority-rule consensus tree is not identical to any of the five trees. Although shown here as if rooted, the trees are considered unrooted in the computation.
A consensus tree is, as we shall see in more detail in Chapter 30, a tree that summarizes a series of trees. Margush and McMorris's majority-rule consensus tree is simply a tree that consists of those groups that occur in a majority of the trees. It may not be obvious that these will form a tree. In fact, they will. If two groups each occur in more than 50% of the trees, then there must be at least one tree that has both of them. If two groups are compatible, then they are either disjoint, or identical, or one must be contained within the other. Suppose that we make up for each group a 0/1 character, which has 1s for each species that is in the group, and 0s otherwise. The compatible groups will then all have compatible characters. The pairwise compatibility theorem that we saw in Chapter 8 then guarantees that all these groups can be placed on the same tree.

The majority-rule consensus tree is found by tabulating all groups that occur on all trees and retaining those that occur on a majority of the trees. When we use it on the bootstrap estimates of the tree, the result is a single tree. All of the groups that appear on it are present in more than 50% of the bootstrap estimates. A simple extension of the majority-rule consensus tree is to note, next to each group, in what fraction of bootstrap replicates it has appeared. We can quickly see which groups have strong support, and which weak support. Figure 20.3 shows five trees and the resulting list of partitions of the species, as well as the majority-rule consensus tree.

The $P$ value for each branch is intended to give an estimate of the amount of support the branch has. As we shall see below, this number turns out to be biased, underestimating the value of $P$ when it is large.

**The multiple-tests problem**

One problem with the use of these $P$ values is that we may not know in advance which group interests us. If we instead look for the most strongly supported group on the tree and then report its value of $P$, we have a "multiple-tests problem" (Felsenstein, 1985b). If there were actually no significant evidence for the existence of any of the groups, then $P$ values on the branches would be drawn from a uniform distribution, with 5% of them expected to fall above 0.95. So one out of every 20 branches of a tree would be expected to reach the "significance" level of 0.95.

One way to correct for this is to use the well-known Bonferroni correction. In this case that simply amounts to dividing the desired tail probability (say 0.05) by the number of tests. Thus if we want to know for which value of $P$ the most significant out of $n$ tests has only a 5% chance of reaching that value, when the null hypothesis (of no significant structure) is true, we should require our groups to attain a support of $P = 1 - 0.05/n$. Thus with (say) 15 groups in a tree, the $P$ value required for significance would be taken to be $1 - 0.05/15 = 0.99666$. This is a conservative procedure and allows for us to find the most significantly supported group out of $n$, even when the support for different groups is not quite independent.
Independence of characters

The most telling criticism of the bootstrap for phylogenies is that the assumptions of independence of the characters may not be met (Felsenstein, 1985b). The easiest way to see what effect this has is to imagine a case in which pairs of characters are identical. In other words, in collecting characters, we have inadvertently collected two characters that are so closely correlated that they are effectively providing the same information about evolution. We have done this so often that each character has, somewhere in our data, an identical partner.

A little consideration will show that the proper method of bootstrapping would be to draw once for each identical pair, as we then have \( n/2 \) independent characters, not \( n \). The proper bootstrapping technique would be to draw \( n/2 \) times, each time drawing one character. If instead we draw \( n \) times, we will be sampling too often, the variation between bootstrap samples will be too small, and the trees they generate will be too similar. There will appear to be more corroborating evidence for groups on the tree than there really is.

Less complete correlation between characters is more realistic. It will cause similar problems — the appearance of too much evidence for groups on the tree. Unfortunately, there is usually no easy way to know how much correlation there is between characters, and thus no easy way to choose the number of characters to draw in a bootstrap sample. In certain cases, such as molecular sequences, one may be able to assume that the correlation of characters occurs mostly between nearby sites in the sequence. For example, we might have correlations that are mostly between sites that are within five nucleotides of each other.

Künsch (1989) has proposed a block bootstrap method that can cope with that correlation. He suggests drawing, not single sites, but blocks of \( B \) sites, the starting position for each block being drawn at random. Instead of drawing \( n \) individual sites, he draws \( n/B \) blocks of \( B \) sites, so that the bootstrap sample ends up consisting of \( n \) sites. Künsch shows that this corrects for autocorrelations along the sequence that are no longer than \( B \) sites. If the distance between correlated sites averages five sites, then \( B = 10 \) would seem to be a good choice. If we are mistaken and there is actually no autocorrelation, block-bootstrapping has the happy property of being a correct method anyway.

Note that in the imaginary example above, where pairs of characters have perfect correlation, if these pairs were adjacent characters, the data set would consist of \( n/2 \) adjacent pairs. One could use Künsch's method with, say, \( B = 4 \) in such a case.

Identical distribution — a problem?

In drawing a statistical sample, one commonly assumes that the draws are independent and identically distributed (i.i.d.). This is also the assumption of the bootstrap. We have seen that nonindependence is a potentially serious difficulty
for the bootstrap, particularly if the dependent characters are not adjacent. Is failure to be identically distributed an equally difficult problem? I don’t think so.

It is evident that the evolutionary processes in different characters (and in different sites in a molecule) can differ substantially. The differences in evolutionary rate from site to site in molecules are one example. Given that, is there any way to use the bootstrap? The approach I have proposed in such cases (Felsenstein, 1985b) is to consider the characters as samples from a larger pool of characters. Suppose that rates are assigned independently to sites in a molecule, so that each site has a rate randomly drawn from a distribution of rates. The characters have randomly assigned rate of evolution, and then the outcome of evolution is the result of a random process running at that rate. To get the data for a character, we draw a rate from the pool of rates, then evolve the character independently at that rate. In that case, the outcomes at the characters are still i.i.d., even though their rates of evolution differ.

In that original paper, I may have created unnecessary difficulties by saying that the bootstrap assumes that “each character is ... a random sample from a distribution of all possible configurations of characters,” and by describing the systematist as sampling from “a pool of different kinds of characters.” Others (Carpenter, 1992; see also Sanderson, 1995) have rejected this argument by disagreeing with the notion that characters are drawn from the universe of all possible characters. Although the notion of there being such a universe is indeed dubious, it is not actually necessary to the argument. All we need to assume is that the characters are drawn independently from some universe of characters, from some pool of characters.

In both molecules and morphology we may have characters that occur in blocks, such as data sets that have 10 skull characters followed by 10 limb characters, or molecules that have a fast region followed by a slow region. The issue that these data sets raise is not identical distribution, but independence. If we could consider successive characters as independently drawn, having a mix of rates of evolution, or a mix of body regions, would not endanger the bootstrap. The existence of these blocks of characters calls into question the assertion of independence, but the heterogeneity of evolutionary processes in the different characters is not the problem.

Invariant characters and resampling methods

The bootstrap and related resampling methods have also been argued to be sensitive to the number of invariant characters included in the data set. Suppose that we are using a method of phylogenetic inference, such as parsimony, that is not affected by the presence of characters that show no variation. Will we get substantially different bootstrap values by omitting the invariant characters from the analysis? It has been repeatedly argued (Faith and Cranston, 1991; Carpenter, 1992; Kluge and Wolf, 1993; Farris et al., 1996; Carpenter, 1996) that the bootstrap
Table 20.1: The probability of a character being omitted from a bootstrap sample, for different numbers of characters ($N$) in the data set.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$(1 - 1/N)^N$</th>
<th>$N$</th>
<th>$(1 - 1/N)^N$</th>
<th>$N$</th>
<th>$(1 - 1/N)^N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>14</td>
<td>0.35434</td>
<td>60</td>
<td>0.36479</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>15</td>
<td>0.35526</td>
<td>70</td>
<td>0.36524</td>
</tr>
<tr>
<td>3</td>
<td>0.29630</td>
<td>16</td>
<td>0.35607</td>
<td>80</td>
<td>0.36557</td>
</tr>
<tr>
<td>4</td>
<td>0.31641</td>
<td>17</td>
<td>0.35679</td>
<td>90</td>
<td>0.36583</td>
</tr>
<tr>
<td>5</td>
<td>0.32768</td>
<td>18</td>
<td>0.35742</td>
<td>100</td>
<td>0.36603</td>
</tr>
<tr>
<td>6</td>
<td>0.33490</td>
<td>19</td>
<td>0.35798</td>
<td>150</td>
<td>0.36665</td>
</tr>
<tr>
<td>7</td>
<td>0.33992</td>
<td>20</td>
<td>0.35849</td>
<td>200</td>
<td>0.36696</td>
</tr>
<tr>
<td>8</td>
<td>0.34361</td>
<td>25</td>
<td>0.36040</td>
<td>250</td>
<td>0.36714</td>
</tr>
<tr>
<td>9</td>
<td>0.34644</td>
<td>30</td>
<td>0.36166</td>
<td>300</td>
<td>0.36727</td>
</tr>
<tr>
<td>10</td>
<td>0.34868</td>
<td>35</td>
<td>0.36256</td>
<td>400</td>
<td>0.36742</td>
</tr>
<tr>
<td>11</td>
<td>0.35049</td>
<td>40</td>
<td>0.36323</td>
<td>500</td>
<td>0.36751</td>
</tr>
<tr>
<td>12</td>
<td>0.35200</td>
<td>45</td>
<td>0.36375</td>
<td>1000</td>
<td>0.36770</td>
</tr>
<tr>
<td>13</td>
<td>0.35326</td>
<td>50</td>
<td>0.36417</td>
<td>$\infty$</td>
<td>0.36788</td>
</tr>
</tbody>
</table>

will give substantially different results without the invariant characters. Harshman (1994) has argued that it will not.

Consider a single character that does show variation in the data set. How often will it appear in the bootstrap replicates? If there are $N$ characters in all, it will be chosen with probability $1/N$ each time a character is sampled. Thus it will be omitted $1 - 1/N$ of the time for each character sampled. The chance that it will be omitted entirely is thus (Harshman, 1994) $(1 - 1/N)^N$.

Adding $M$ invariant characters to a data set changes this probability by increasing the value of $N$. Harshman argues that this quantity is very close to being constant at $e^{-1} = 0.36788$, no matter what the value of $M$. Farris et al. (1996) argue that it is not constant, that its complement (the probability of the character being included) “decreases as $N$ increases.” Table 20.1 shows the probabilities of the character being omitted.

The values do increase (and the probabilities of inclusion decrease), but not by much: They reach 90% of their ultimate value at $N = 6$, and 99% of the ultimate value at about $N = 50$. We can conclude, with Harshman, that the inclusion or exclusion of invariant characters will have little effect on the support given any particular group by the bootstrap method. The delete-half jackknife will behave similarly.

Of course, if the method for inferring phylogenies assumes that all characters are present (as do distance and likelihood methods), then we cannot drop invariant characters without doing serious violence to the trees.
Biases in bootstrap and jackknife probabilities

For years after the introduction of the bootstrap method for phylogenies, people had complained that the \( P \) values that the bootstrap method provided seemed too pessimistic. When they were noticeably lower than 95%, there still seemed to be a very high chance that the groups were real. Zharkikh and Li (1992; Li and Zharkikh, 1994) carefully examined the statistical properties of this inference and showed that the support was indeed underestimated. Hillis and Bull (1993) carried out a large simulation study that reached the same conclusion. They argued that a \( P \) value as small as 70% might indicate a significantly supported group. Felsenstein and Kishino (1993) have agreed that the bias is present, but they argued that it is not due to the bootstrap sampling itself, but instead to the use of a \( P \) value to describe the presence or absence of particular clades. Efron, Halloran, and Holmes (1996) argued that there was not always a bias downwards; they are correct, but for high values of \( P \) the bias is almost entirely in that direction. Newton (1996) has verified the validity of the bootstrap for discrete entities such as tree topologies, and has also verified that there is this bias.

\( P \) values in a simple normal case

To show that this bias is not due to the bootstrap, we argued that it would appear even in cases where there was no bootstrapping. For example, suppose that we draw \( n \) points from a normal distribution whose standard deviation is known to be 1, but whose mean is unknown. We are interested in whether the mean is positive or negative. This is analogous to asking whether a branch is present or absent, with the value of the mean playing the same role as the branch length. Our estimate of the mean will be the empirical mean of the sample, \( \bar{x} \).

To obtain a level of significance for the proposition that the true mean is positive, we consider that the sample mean is normally distributed around the true mean with variance \( 1/n \). The conventional way of constructing \( P \) values is to use pivotal statistics. Thus we have the difference between the true mean, \( \mu \), and the sample mean \( \bar{x} \), which is \( \bar{x} - \mu \). That difference has a normal distribution with mean 0 and variance \( 1/n \). It follows that when we multiply it by \( \sqrt{n} \) it will become a quantity with mean 0 and variance 1. The probability that this quantity is greater than some particular value is then easily computed from tables of the normal distribution. We can then say, for example, that there is a 95% probability that \( \sqrt{n} (\bar{x} - \mu) \) is greater than 1.64485. This can be turned into a statement assigning a level of significance to the statement that \( \mu > 0 \). For example, if \( n = 10 \) and \( \bar{x} = 0.7 \), we know that \( \sqrt{10} (0.7 - \mu) \) has a normal distribution with mean 0 and variance 1. The probability that \( \mu < 0 \) is then the probability that a standard normal deviate lies below \( 3.162 \times (-0.7) = -2.214 \), which is about 0.014. Then the probability that \( \mu > 0 \) is thus approximately 0.986.

We will have to return and ask what this really means. It seems entirely too neat (and so it is). But for the moment it tells us how to assign a \( P \) value to the
Figure 20.4: An example of assigning values of $P$ to regions of a space that resemble tree topologies. We draw a sample of points from a true distribution (dashed curve) and there is a resulting distribution of the mean of $n$ such points (density function with darkest line). The two other density functions show what we might infer this density function to be if the mean came out a bit closer to, or a bit farther away from 0. In each case the $P$ value assigned is given by the shaded area of the curve.

The correct $P$ value to assign is the tail area of the true distribution of $\bar{x}$, which tells us the probability that our samples will get the true “tree topology.” The actual $P$ values vary around this, and it is immediately apparent that they do not vary symmetrically. When $\bar{x}$ is too close to 0, they drop substantially. When it is too far from 0, in an event that is equally likely to occur, the $P$ rises by a much smaller amount.

The result is that there is a bias in $P$. When $P$ should be (say) 0.95, the value we get is on average smaller than 0.95, leading to statements that are on average too conservative. Figure 20.5 shows the average $P$ values as a function of the true
Figure 20.5: The expected value of $P$ for the hypothesis that $\mu > 0$ in the case of $n$ points drawn from a normal distribution with expectation $\mu$ and variance 1 (as in Figure 20.4). The expectation of $P$ is plotted as a function of the true probability that a sample will have $\bar{x} > 0$.

$P$ values, which are easily computed for this example (Felsenstein and Kishino, 1993). The bias of $P$ is apparent. It is always toward 0.5, which, for the large values we are interested in, means that the $P$'s are on average too conservative.

In Figure 20.4 we can also see that when the "true" value is $P$, the estimate $P_e$ will be greater than $P$ half of the time, and less than $P$ the other half of the time. It is less obvious, but also true, that the estimate $P_e$ will be greater than 0.5 a fraction $P$ of the time. Thus when the true $P = 0.95$, the estimated $P$ will exceed 0.95 half of the time, and the fraction of times that $P_e$ will exceed 0.5 is 95%.

One of the sources of the conservatism of the estimated $P$ values is that we are taking statements about the "branch length" $\mu$ and reducing them to statements only about the "tree topologies" $\mu > 0$ and $\mu < 0$. If the observed mean turns out to be above $1.95996/\sqrt{n}$ (the one-tailed 95% point of a normal distribution), we will conclude that the confidence set is entirely of "topology" I. When it is below $-1.95996/\sqrt{n}$, we conclude the opposite, that it is entirely of "topology" II. Anywhere in between, we will conclude that both "topologies" are possible.

If the true value of $\mu$ were (say) infinitesimally less than 0, so that the "topology" was II, but just barely so, we would draw the wrong conclusion 5% of the time, as that is how often we would get an observed mean that exceeded $1.95996/\sqrt{n}$. The other 95% of the time the confidence set would include the cor-
rect “topology.” For any more negative value of \( \mu \), the probability of type I error (falsely rejecting the true “topology”) is less than 5%, often considerably so. For example, when \( n = 10 \) and \( \mu = -0.1 \), the value \( 1.95996/\sqrt{10} = 0.61979 \) is 2.27619 standard deviations away from that true mean. Thus we get the false conclusion that the “topology” is only about 0.0114 of the time, as this is the fraction of a normal distribution that lies beyond 2.27619.

These results are true for the analogy of tree topologies with regions of positive and negative values of a normally distributed quantity. Will similar behavior be seen for actual tree topologies? This is not known, but I suspect that topologies will behave very similarly.

This analogy leads us to one interpretation of the bootstrap \( P \) value. If we see a group that occurs a fraction \( P \) of the time, we can say that the probability that it would have obtained this much support if it were not actually present on the true tree is less than \( 1 - P \). Thus a group that obtains a \( P \) value of precisely 95% will be expected to obtain that much support, when it is not actually present, less than 5% of the time. We must, however, note that the proof of this conservative interpretation has not yet been made for the case of phylogenies.

**Methods of reducing the bias**

The bias of the \( P \) value becomes even greater when consider that we are in a space of trees and consider the multiple topologies near each tree. As we will see, the effect is to increase the bias. Four methods have been proposed to correct for this bias. I will describe each briefly, and then suggest some connections between them.

- **The complete and partial bootstrap.** Zharkikh and Li (1995) developed a method which at the time seemed strange. It looks much less strange now that it has been joined by other methods and the connections between them become more apparent. Zharkikh and Li considered a case where there were \( K \) different character patterns, each backing a different tree topology. Using normal approximations and simulations, they showed that the bias of the bootstrap \( P \) value grew greater as \( K \) got much larger than 2. They went on to derive the complete and partial bootstrap method to correct these \( P \) values. We do not know what the relevant value of \( K \) is for a space of tree topologies. But they noted that for the case of \( K \) classes, if we do two bootstrap analyses with different numbers of characters resampled, we can estimate the effective values of \( K \) and of the probability of the correct class, and then use it to correct the bias. Suppose that we draw a regular bootstrap sample and obtain \( P = P^* \). We also do partial bootstraps, in which we sample only \( 1/r \) as many characters (thus, if \( r = 3 \), we resample a number of times only one-third the number of characters). Call the fraction of these smaller resamplings that support that particular outcome \( P^*_r \). Zharkikh and Li then were able to compute from the values of \( P^* \) and \( P^*_r \) what was the effective value of \( K \), and use that to correct the bootstrap \( P \) value.
• The method of Efron, Halloran, and Holmes. Efron, Halloran, and Holmes (1996) applied a correction due to Efron (1987) to get a less biased P value for presence of a group in a phylogeny. They first bootstrap the data and infer trees. I have noted above that bootstrapping can be regarded as a reweighting of characters, where each of the original characters has a weight corresponding to the number of times it occurs. Thus if character \( i \) occurs \( n_i \) times, this would be the same as having it have weight \( n_i \). They now take the samples that do not show the particular group, such as \{Human, Chimp\}. For each of these they try to adjust the weights back toward equality, so as to arrive at a set of weights that results in the group just barely being absent. One searches for the fraction \( f \) that determines weights \( f + (1 - f)n_i \), such that these weights just barely result in the absence of the group. Efron, Halloran, and Holmes point out that this can be done by a simple “line search.” The data set with these weights is a least favorable case, one that lacks the group but comes as close as possible to the original data set. They now bootstrap from these reweighted data sets. If the weights are \( w_i \), the bootstrap draws character \( i \) with probability \( w_i \). Analyzing this second level of bootstrap samples, they see what fraction of the resulting trees contain the group. After computing a constant \( a \) from the weights for each of these reweighted data sets, they then use a formula from Efron (1987) to calculate a bias-corrected P value.

• The iterated bootstrap. Rodrigo (1993) adapted methods invented by P. Hall and R. Beran in the statistical literature to propose the iterated bootstrap. He uses no less than three levels of bootstrapping. First one takes the usual \( R \) bootstrap replicates and estimates the tree for each. Then for each of these bootstrap sampled data sets, one bootstraps \( R \) more times from it, so that one has done \( R + R^2 \) bootstrap samples in all. Not content with this, one goes one more level, to make a triple bootstrap with a total of \( R + R^2 + R^3 \) replicates. We assume that our interest is in some particular group (such as \{Human, Chimp\}), and we want to discover what fraction of times \( P \) it should appear in the bootstrap estimates to make its appearance give us 95% confidence in its existence.

We would ideally like to know the true tree, sample more data sets generated on it, and see how often we rejected the group when it was present on the true tree. This we cannot do: If we knew the true tree, we would not even bother to ask the remaining questions. The iterated bootstrap takes the \( R \) bootstrap estimates of the tree as true, and for each takes the \( R^2 \) second-level bootstrap samples to approximate the variation of data generated on such trees. Then the third level of sampling is used to find out, for each of these \( R^2 \) data sets, whether the group in question would be judged to have significant support. This is done for the first-level trees that have the group and for the first-level trees that do not. These are used to approximate the proba-
bility that a group that is not present will be significantly supported, and the probabilities that a group that is present will be significantly supported.

- **The AU method of Shimodaira.** Shimodaira (2002) has developed a method similar to Zharkikh and Li’s complete and partial bootstrap. It uses a series of bootstraps of different sizes. One might be the original bootstrap, another might sample $n/2$ sites, and another $2n$ sites (which is perfectly possible since sampling is with replacement). By fitting curves through the resulting $P$ values, he obtains constants needed for a correction formula. Shimodaira and Hasegawa (2001) have described a computer program to do this.

The correction formulas for three of the methods look generically similar, which suggests that the methods are related. They have much in common. The first two explore the shape of the region of data space that lead to inferring the group. The partial bootstrap (used in the Zharkikh and Li method and in the AU method) has us spread out more widely from the original data set and see what this does to the probability of inferring the presence of the group. The Efron-Halloran-Holmes (EHH) method moves to the nearest edge of the region and uses the bootstrap to ask about the geometry of the region there. They argue that they are in effect asking about the convexity of the region in that neighborhood. Shimodaira discusses the matter in more detail and points out the close relationship of his method with these two methods. It is less easy to see that the iterated bootstrap also does something similar, as it works more empirically without any explicit geometry.
The methods differ in computational effort. The iterated bootstrap can be quite tiresome, as it replaces bootstrap \( R \) replicates with more then \( R^3 \) replicates. This would replace 100 replicates by more than a million replicates. The method of Efron, Halloran, and Holmes takes a fraction of the bootstrap replicates, reweights their characters, and then resamples from these. This will be tedious but not nearly as burdensome as the iterated bootstrap. The complete and partial bootstrap method and the AU method are the least difficult because they can be carried out with as few as two bootstrap samplings. However, those samplings may need a large number of replicates to obtain sufficiently accurate \( P \) values. Shimodaira presents computer simulation results comparing the ZL, AU, and EHH methods, and finds that AU is most accurate.

We can make tables to carry out both of these methods with two bootstraps. Suppose that we have a complete bootstrap plus a partial bootstrap that samples half as many characters. Call their observed bootstrap \( P \) values \( P^* \) and \( P_{r^*} \), respectively. Table 20.2 shows for each method which values of \( P^* \) are small enough to allow the bias-corrected \( P \) to reach 0.95 for a number of different \( P \) values for the complete bootstrap.

**The drug testing analogy**

In Hillis and Bull’s (1993) simulations, they asked what fraction of the time a group that had 95% bootstrap support would be on the true tree. They found that groups that had as little as 70% support had a 95% chance of being true. This was the outcome of a simulation in which they took randomly branching trees and evolved characters along them.

Will this prove to be a general result? If so, then we might hope for general rules allowing us to correct the \( P \) values and interpret the result as a probability that the group is correct. The following Bayesian analogy shows that there is some reason for doubting this. Suppose that we are carrying out product tests for a pharmaceutical company, testing whether their drugs cure a particular disease. We do a blind test of the proposition that the drug is ineffective, and come up with a tail probability \( \alpha \) for the test. Some of the time we reject this null hypothesis. Consider a group of proposed drugs that have each achieved \( \alpha = 0.05 \). What fraction of them actually work?

It depends heavily on who selected the drugs. They are submitted to us by the drug development branch of the company. If that branch is highly competent, they will submit to us mostly drugs that work. In that case many of them will reach the \( \alpha = 0.05 \) threshold, and the probability that a drug that reaches \( \alpha = 0.05 \) actually works is then very high, probably much higher than 0.95. On the other hand, if the drug development branch is not competent, then the drugs they submit for testing will mostly be ineffective. Few drugs will reach the 0.05 threshold, and when one does, it will have a small chance of being one that actually works, being more likely to be one of the 1 drugs in 20 that accidentally appears to work.
Hillis and Bull (1993) had, in effect, a fairly competent drug development laboratory. They used computer simulation on randomly branching trees. If there is a moderate amount of evolution on the branches of the tree, and a large number of characters, the groups recovered will tend to have a large probability of being correct. If, however, there is too much change between nodes on the tree, the groups recovered will reflect mostly random noise, and have a good chance of being incorrect.

We can use the normal distribution analogy to show this phenomenon. Suppose that \( \mu \) itself is drawn from a normal distribution with mean 0 and variance \( \sigma^2 \). We know that, for \( n \) characters, a group reaches \( P = 0.95 \) when its sample mean \( \bar{x} \) is \( 1.96996/\sqrt{n} \). If we take \( n \) data points from a normal distribution with variance 1, whose mean is itself normally distributed with mean 0 and variance \( \sigma^2 \), that mean, \( \bar{x} \), will come from a normal distribution with mean 0 and variance \( \sigma^2 + 1/n \). We can now ask about the conditional distribution of the true \( \mu \) given the observed \( \bar{x} \). This too is normal. It has mean \( b_{\mu,\bar{x}} \bar{x} \), where \( b_{\mu,\bar{x}} \) is the regression of \( \mu \) on \( \bar{x} \). That regression is the fraction of the total variance \( \sigma^2 + 1/n \) which comes from the variation of \( \mu \), namely

\[
    b_{\mu,\bar{x}} = \frac{\sigma^2}{\sigma^2 + \frac{1}{n}} \quad (20.1)
\]

The variance of \( \mu \) given \( \bar{x} \) is also easy to obtain. It is the residual variance in \( \mu \) after the variance due to regression is taken out, which can be calculated to be

\[
    \sigma^2 - b_{\mu,\bar{x}}^2 \left( \sigma^2 + \frac{1}{n} \right) \quad (20.2)
\]

Using equation 20.1, this variance is easily shown to be

\[
    \text{Var} (\mu | \bar{x}) = \frac{\sigma^2}{n\sigma^2 + 1} \quad (20.3)
\]

Thus given a group that has significance level \( P \), we can calculate the probability that it truly has \( \mu > 0 \). All we need to do is (1) find the standard normal deviate that has area \( P \) below it, (2) multiply this by \( \sqrt{\sigma^2 + 1/n} \) to get the corresponding value of \( \bar{x} \), (3) multiply that by the regression coefficient \( \sigma^2 / (\sigma^2 + \frac{1}{n}) \) to get the mean of \( \mu \), (4) calculate how many standard deviations this is from 0 when the variance is given by equation 20.2, and (5) work out what fraction of that conditional distribution of \( \mu \)'s lies above that point. Note that 0 lies in the left tail of this distribution of \( \mu \)'s, and thus we are asking about the area above that point.

This has been done by Felsenstein and Kishino (1993). Figure 20.6 shows the results, with the probability that \( \mu > 0 \) plotted against \( P \). The result depends on the value of \( n\sigma^2 \), and these values are indicated next to the curves. When \( n\sigma^2 = 0.1 \), in effect there is very little genuine signal (the drug development group...
Figure 20.6: The probability that $\mu > 0$ when we draw $n$ points from a normal distribution whose expectation $\mu$ is itself normally distributed with mean 0 and variance $\sigma^2$. The probability is plotted as a function of the $P$ value for the observed mean. The value of $n\sigma^2$ is shown next to each curve.

is sending drugs that are generally ineffective). Even when a test reaches $P = 0.95$, the probability is not much greater than 50% that the true mean is above 0. When $n\sigma^2$ is 1, the curve is nearly a straight line, and when a test reaches $P = 0.95$, it has a bit more than 95% chance that the true mean is above 0. Hillis and Bull's (1993) results looked more like the case $n\sigma^2 = 2$, as they found that when $P = 0.70$, the group appeared on the true tree about 95% of the time.

These results are for the normal distribution analogy. What use can a user of the bootstrap make of them? Until further simulation testing on phylogenies is done, one has to be cautious. We do not know whether Hillis and Bull's rule of thumb is general. We do not know whether other cases are similar in the parameters that correspond to $n\sigma^2$. Note that with more information (larger $n$) the bootstrap becomes more conservative. One way to get a feel (but no more than that) for the conservatism of the bootstrap would be to look at all the $P$ values on the tree.
If they are all large, this indicates that \( n \) is large, and we may then cautiously conclude that \( P \) values much less than 95% may indicate groups that have a high probability of being true. But if the \( P \) values are mostly small, then \( n \) is not large and we must be much more cautious in concluding that they indicate that a group is true.

Berry and Gascuel (1996) have argued that if correctness of trees is judged by the symmetric difference metric (which will be explained in Chapter 30), and if we could Type I and Type II errors as equally serious, the best value of \( P \) to use to resolve the tree partially would be \( P = 0.5 \). Their argument relies on a particular form of the relationship between the measured \( P \) value and the probability of the grouping being correct, one which makes this probability 0.5 when \( P = 0.5 \). It seems unlikely that this is true in general, so that their proposed rule needs further examination.

**Alternatives to \( P \) values**

Another difficulty with \( P \) values on groups is that one "rogue" species that is of uncertain placement can disrupt the signal in a majority-rule consensus tree. If the group ABCDEF occurs in most trees, but half of the time with species G in it and half of the time without, the majority-rule consensus tree may not contain either ABCDEF or ABCDEFG. The majority-rule method does not give a group credit for a partial appearance, or for appearance only in a larger group. Sanderson (1989) has suggested coping with this by setting a number \( n \) of extra individuals allowed into a group. Thus, if \( n \) is 2, we note that ABCDEF is present whenever a group containing those species and no more than 2 others is present. In the example above, ABCDEF would be given high support when \( n = 1 \), as then ABCDEFG would count towards it being present.

Wilkinson (1996) proposed another method: computing a reduced majority-rule consensus tree which shows trees of groups that are present among the bootstrap estimates of the trees, when we drop various species from consideration. Thus, dropping species G, we find ABCDEF present a large fraction of the time. He did not present efficient algorithms for finding the set of reduced majority-rule consensus trees. He notes that they require us to specify the desired tradeoff between number of species dropped and strength of support for groups. Algorithms to find these trees efficiently are still lacking.

In both cases some of the problems from noise are reduced by asking a somewhat looser question. Computational issues aside, the question that must be faced is whether this looser question is meaningful enough. Is it helpful to know that the group \{Human, Chimp\} occurs often if some additional species are allowed in the group, if the broader group turns out to be \{Human, Chimp, Mouse\}?

Brown (1994a) suggests other questions: Does a group appear significantly more frequently than another, and does a group appear significantly more often than 50%? I cannot see that these are useful: With enough bootstrap replicates a group that appears 51% of the time will be declared to appear significantly more
often than 50%. But does this mean that its appearance on the true tree is supported? I suspect not.

**Probabilities of trees**

An alternative to the puzzle of how to describe support for groups is to simply take the distribution of trees and measure support for the different tree topologies. If we have a modest number of species we may be able to look at all possible trees. With 5 species and unrooted trees, there are 15 bifurcating tree topologies, and we can count how often each of them occurs among the bootstrap estimates of the topology. One way of constructing a confidence interval on trees is then to take the most frequent topologies until their probabilities add up to at least 95%. As the number of species increases, it will be less and less practical to do this. The number of possible phylogenies increases greatly, and it will soon become rare that two bootstrap replicates will lead us to estimate the same tree topology. We then end up with two classes of tree topologies—those that occurred once, and those that did not occur. We might order the ones that occurred once according to their goodness-of-fit to the original data (as judged by likelihood, parsimony, or whatever criterion we are using). The real problem is that we are then not concentrating our attention on the trees that contain a group of interest, so that we lose power in evaluating such a group.

Tree probabilities estimated from a bootstrap are used in Lake’s (1995) “bootstrapper’s gambit” method. There each bootstrap sample has its quartets analyzed, and if these all are compatible, a tree is constructed from them. When the tree probabilities are calculated, their interpretation is marred by the omission of all bootstrap samples that have incompatibilities among their quartets. Lake’s tree probabilities must therefore be regarded as upper limits on the actual values.

**Using tree distances**

In Chapter 30 distance measures between trees will be described, in particular the symmetric difference metric. Penny and Hendy (1985, 1986) used this difference, together with the jackknife, to discover how far from the true tree we are. They randomly sampled a fraction of all characters, and constructed a tree from this resampled data. They calculated the mean distance between the trees from different samplings. They could show that, as the fraction of characters that were sampled increased, the trees became closer to each other. Plotting the decline of distance between trees against the number of characters sampled allowed them to infer how much sequence data was necessary to infer the true tree accurately.

Miller (2003) has used a similar plot (although using distance from a reference tree rather than distance between different sampled data sets). Like Penny and Hendy, his interest is in distances between trees, in order to understand the accuracy of the whole tree.
Jackknifing species
Early on, Lanyon (1985) suggested using a jackknife across species, removing one species at a time from the tree to see what effect this had on the estimate of the relationships of the remaining species. It is not easy to see what statistical meaning this jackknifing of species will have. Species are not independent and identically distributed — they come to us on some phylogeny, where they are highly clustered. This has been a major barrier to any attempt to make a statistical interpretation of jackknifing or bootstrapping species instead of characters.

Parametric bootstrapping
In the bootstrap, the resampling of the data set is intended to mimic the variability that we would get if we could sample more data sets from the underlying true distribution. In effect, that would be what we would get if we could simulate data sets on the true tree using the true model. The data sets we get from bootstrapping would be similar in the kinds of variability they contained. As we have seen in the discussion of biases, the trees they yield vary around the estimate that the original data set gives rather than around the true tree.

On the assumption that our estimate of the tree is somewhere near the true tree and that our model is somewhere near the true model, we could also imagine using our estimate and making new data sets on it by computer simulation. We would hope that these data sets also contain the same kinds of variability as would

---

**Figure 20.7:** The parametric bootstrap. The data sets are obtained by simulation on our best estimate of the tree rather than by resampling columns of the original data matrix.
data sets from the true tree. The data sets could be treated in much the same way as are bootstrap samples. This method is the parametric bootstrap. The technique was introduced by Efron (1985). It was introduced for phylogenies by a number of people (Felsenstein, 1988b; Goldman, 1993; Adell and Dopazo, 1994; Huelsenbeck, Hillis, and Jones, 1996).

The closeness of the relationship between parametric bootstrapping and the ordinary bootstrap has led to the latter being referred to as the nonparametric bootstrap. With a single variable this is particularly apparent. Sampling from the original data is the same as sampling from an empirical histogram of data points. This histogram is regarded as an estimate, hopefully a close one, of the original distribution from which the data were drawn. Parametric bootstrapping replaces this histogram with a distribution from a parametric family, with the parameters being those that would be inferred from the data.

Figure 20.7 diagrams the process of using the parametric bootstrap with $R$ replicates:

1. A single best estimate of the tree is made from the data set.
2. $R$ (in this case, 100) computer simulations are then used to produce $R$ data sets of the same size from this tree.
3. Each of these simulated data sets is used to infer the tree, using the same method used on the original data set.
4. The resulting trees are then analyzed in the same way as in the ordinary (nonparametric) bootstrap, such as by making a majority-rule consensus tree and $P$ values for branches in the tree.

**Advantages and disadvantages of the parametric bootstrap**

Parametric bootstrapping can be used as a general replacement for nonparametric bootstrapping. For small data sets, it will have the advantage that it can sample from the desired distribution, even when sampling columns of the data matrix might leave many kinds of variation in the data unrepresented. The main concern is its close reliance on the correctness of the statistical model of evolution. When the model is correct, the type of variation that we will get between different bootstrap sample data sets will closely mimic the type of variation that we will get between the simulated data sets. It will not matter much whether we use parametric or nonparametric bootstrapping. But when the model is not correct, they will behave differently. The sampling of columns of the data matrix in ordinary nonparametric bootstrapping will reflect the variation in the correct model, while the simulation in parametric bootstrapping will reflect the variation in our incorrect model. In this situation, the ordinary (nonparametric) bootstrap will have the advantage. The more trust we have in the adequacy of our model, the more we will be willing to use instead the parametric bootstrap.
Permutation tests

An alternative to resampling is to reorder one's data. Permutation tests are standard methods in nonparametric statistics. For example, if we have two samples, one with 34 points and one with 43 points, that are supposed to be drawn independently from the same distribution, we can do a nonparametric version of a t-test by computing the difference in their means. Rather than assume we know the distribution from which they came, we can simply reshuffle the points many times. Suppose that we take all 77 values and shuffle them into a random order. Take the first 34 as being in sample 1, the second 43 as being in sample 2. Compute the difference in their means.

If we continue shuffling into random orders, and each time compute the difference of means, we get a large sample from the distribution of means under the null hypothesis that the two samples are from the same distribution. If we draw (say) 999 such samples, we can take these differences of means, and consider also the actual difference of means. Of these 1,000 numbers, if the actual difference lies in the top 25 or the bottom 25, we can reject the null hypothesis with $\alpha = 0.05$. Under the null hypothesis, all 1,000 values are from the same distribution, and the probability of being in these tails is 0.05.

Notice that the samples are not precisely from the full distribution because they always involve the same 67 numbers. Notice also that there are only a finite number of possible outcomes. There are only $77!/(34!43!)$ possible outcomes, but this is a satisfyingly large number, over $8.1 \times 10^{21}$. The power of the test is also dependent on the intelligent choice of a statistic. If the underlying distribution is one that generates samples whose means are dominated by a few extreme values, this test would not be particularly sensible.

A number of permutation strategies have been suggested:

**Permuting species within characters**

Archie (1989) and Faith and Cranston (1991) have suggested a permutation test for the presence of taxonomic structure in a data set. It is often called the permutation tail probability test (PTP). They take each character (column) in the data matrix and shuffle its values, reassigning them to species at random. All of the columns are shuffled independently of each other. The hope is that this will produce data sets that have no phylogeny but have numbers and distributions of states that are typical of the data. The distribution of goodness-of-fit measures such as likelihood or parsimony score among these permuted data sets are then compared to the value from the original data. If the actual value lies far enough into the tail (in the direction of higher likelihood or lower parsimony score), then there is significant taxonomic structure in the data. Källersjö et al. (1992) suggest some approximate strategies for more rapidly sampling trees and approximately computing the tail probability, based on Chebyshev's inequality and an exponential approximation to the distribution of tree lengths.
There are two difficulties with the PTP test. One is that structure may be detected for relatively trivial reasons. Suppose that two species are sibling species and that these are nearly identical. This may be enough to cause the test to be significant. It is true that there is then relatedness among species being detected, but it is only this rather obvious relationship between sibling species, and it does not mean that other larger-scale relationships are being detected. For reasons similar to this, Slowinski and Crother (1998) argue that the PTP test too readily detects significant structure. Simulation tests of the PTP method have disagreed whether or not its probability of type I error is too high (Peres-Neto and Marques, 2000; Wilkinson et al., 2002).

A second, and more serious problem was pointed out by Thorne (in Swofford et al., 1996). A tree with only a single internal node, with all lineages branching from it in a great multifurcation, can show significant structure in the test, if the branches are of substantially unequal lengths. An example of such a case is given in that paper. One possible response to this is that such a case does have structure. If one lineage is much longer than the others, and if the tree is unrooted and we regard it as the outgroup, then the other species can be regarded as forming a group distinct from the outgroup. Källersjö et al. (1992) have suggested that, rather than using parsimony score to characterize the degree of monophyly, we use a total support criterion, which is the sum over all branches of the Bremer support values. This will be zero if there is no unambiguous support for any monophyletic group. Farris et al. (1994a) gave an example data set where there was no such unambiguous support for any monophyletic group, but where the PTP test using parsimony score is significant. The example shows definite structure — for example, placing the species in a linear order — so that it is possible to argue that the behavior of the PTP test is appropriate. There has been debate back and forth over these examples (Carpenter, 1992; Faith, 1992; Källersjö et al., 1992; Trueman, 1993; Faith and Ballard, 1994; Farris et al., 1994a; Farris, 1995; Trueman, 1996; Carpenter, Goloboff, and Farris, 1998). The debate revolves around what the null hypothesis and alternative hypotheses of the PTP test really are. Some of these concerns have been raised on philosophical grounds (Goloboff, 1991; Bryant, 1992), but the matter will be more readily resolved in a statistical context. This needs more examination, so that we can understand what are the assumptions and behaviors of the test.

In an effort to concentrate the test’s attention on hierarchical structure, Alroy (1994) has suggested using the PTP permutation strategy but computing different statistics, based on the number of pairs of characters that are compatible.

A variation of the PTP test is suggested by Faith and Cranston (1991). The tree topology is held constant and the data permuted, evaluating each permutation on that topology. This is held to test whether that tree has support greater than random. As this way the tree cannot adapt to the data, the test is quite likely to reject randomness. Brown (1994a) suggests using the permutation while examining whether a particular group appears significantly more often among bootstrap
estimates than among estimates from the permuted data. As the presence of almost any internal structure within the group can cause that to happen, this seems unlikely to be a useful question.

**Topology-dependent permutation of species.** Faith (1991) developed a version of this permutation method designed to test whether a specific branch, which divides the species into two groups, is supported. He permutes the data as above, but instead of computing the total parsimony score for each data set, he computes the Bremer support for a given split of the species into two groups. For each data set we must compute the difference between the best tree and the best tree that does not contain this split. Faith argues for randomization that does not include the outgroup species. Swofford et al. (1996) disagree and prefer randomization of each character over all species. Faith’s randomization test, the topology-dependent permutation tail probability (T-PTP) test, is designed to test whether there is nonrandom support for that particular split.

There has also been some uncertainty as to what should be done with the outgroups in the randomization process. Trueman (1996) argued that exclusion of the outgroup species from the randomization process was appropriate, and suggested ways of making the test more conservative.

As with the PTP test, there has been much discussion of the usefulness and validity of this test (Faith, 1992; Trueman, 1993; Farris et al., 1994a; Faith and Ballard, 1994; Trueman, 1996; Faith and Trueman, 1996; Carpenter, Goloboff, and Farris, 1998). I note here one particular criticism. Swofford et al. (1996) concentrated on whether other, irrelevant structure in the data could cause the test to reject randomness too often. (Farris, 1995 had earlier made an equivalent suggestion.) Swofford et al. simulated evolution on a tree of topology $(O, (A, B, (C, D)))$, and found that the group $(B, C, D)$ was supported too often. The presence of group $(C, D)$ thus made the randomization procedure inappropriate, as it often broke up this group. The T-PTP test may thus have a null hypothesis of no structure anywhere in the tree, which dilutes its focus on the monophyly of the one group. Faith and Trueman (1996) have argued that this criticism is invalid, being based on the wrong choice of null hypothesis.

**Permuting characters**

Many phylogeny inference methods are insensitive to the order of the characters (the exception is likelihood or Bayesian methods that allow for autocorrelation of rates among sites). It might thus seem uninteresting to permute the order of the characters in a data set. But when there are two data sets, we might wish to know whether they are inferring noticeably different trees. If the data sets have, respectively, $n_1$ and $n_2$ characters, this can be addressed by a permutation test. We take all $n_1 + n_2$ characters, and allocate them randomly into two data sets of size $n_1$ and $n_2$. This is most easily done by permuting the order of the $n_1 + n_2$ characters, and taking the first $n_1$ to be the first data set and the second $n_2$ to be the second data set.
The test is carried out by doing this $R$ times, and measuring for all of these replications some aspect of the difference between the phylogenies inferred from the two data sets. We add the original data set into the picture and see whether the difference between their phylogenies is in the top 5% of these $R+1$ numbers. Permutation tests like this are standard in statistics; in systematics they go back at least to the paper of Rohlf (1965), who did not compare phylogenies but measured the correlation between distances inferred from both data sets. Penny and Hendy (1985) used random divisions of a data set into halves to measure the average distance between the resulting trees, and from that get an idea of how accurately the tree was being estimated. The permutation test of whether the trees from two data sets are significantly different was introduced by Farris et al. (1994b) as the incongruence length difference (ILD) test, and independently by Swofford (1995) as the partition homogeneity test. It is most often known by the former name.

For the ILD family of tests, one computes for each replicate (and for the original two data sets) a measure of the extent to which the two data sets result in different trees. This can be done for parsimony, distance matrix, or likelihood methods. For parsimony, suppose that $T(D)$ is the tree estimate from data set $D$, and $N(D, T(D))$ is the number of changes required to evolve data set $D$ on that tree. If the data sets are $D_1$ and $D_2$, and if when combined they are the larger data set $D$, then the suggestion of Farris et al. (1994b; Farris et al., 1995b) is to use the measure of Mickevich and Farris (1981), which is $N(D, T(D)) - N(D_1, T(D_1)) - N(D_2, T(D_2))$, a number that cannot be negative. (I leave it to the reader to discover why.) Swofford (1995) notes that the first term is unnecessary as it is the same in all permutations of a data set.

There are other possible measures. For example, one could use a tree distance (for which see Chapter 30) to measure how dissimilar the two trees are. Generalizations using the criteria for distance or likelihood methods are also straightforward, as long as one takes into account that higher is better in likelihood. In any of these cases one tests whether the measure of difference in outcome for the actual data sets is in the top 5% of the distribution, where the other $R$ replicates are generated by permutation. If it is significantly extreme, this is an indication that the two data sets have significantly different signal.

ILD tests have been fairly widely used to analyze real data sets. The ambiguity in these permutation tests is exactly what a significant result implies. Trees can be different in topology and/or in branch length. Simulations by Dolphin et al. (2000), Dowton and Austin (2002), Darlu and Lecointre (2002), and Barker and Lutzoni (2002) found that inequalities of rates of evolution in different data sets, using the same tree, could cause an elevated rate of rejection of the null hypothesis. This suggests caution in concluding that two data sets imply different trees.

**Skewness of tree length distribution**

A technique that is not really a permutation test, but which should be discussed along with them, is the skewness test of Hillis (1991; see also Fitch, 1979, 1984),
which is discussed more extensively by Huelsenbeck (1991). This looks at the numbers of changes on all possible tree topologies, using parsimony. There is judged to be phylogenetic signal in the data if the distribution is significantly skewed. The rationale for this is that a few trees of much lower score than the others will create negative skewness. It can be computationally burdensome to examine all possible topologies, when the number of species is not small. The burden can be largely avoided by instead sampling randomly from the distribution of all possible topologies.

This method has been criticized by Källersjö et al. (1992), who gave a data set on which it did not behave properly. The fact that skewness is affected by all parts of the tree distribution, and does not concentrate its attention on the better trees, means that it may be of limited power in detecting phylogenetic signal.