RH: POINTS OF VIEW

Is There Something Wrong with the Bootstrap on Phylogenies? A Reply to Hillis and Bull

Joseph Felsenstein[1] and Hirohisa Kishino[2]

[1]Department of Genetics SK-50, University of Washington,
Seattle, Washington 98195

[2]Ocean Research Institute, University of Tokyo,
1-15-1 Minami-dai, Nakano-ku, Tokyo 164, Japan

Send proofs to:
Joe Felsenstein
Department of Genetics SK-50
University of Washington
Seattle, WA 98195

(206)-543-0150
joe@genetics.washington.edu

Keywords:
bootstrap
phylogenies
hypothesis testing
computer simulation
statistics

In a paper in this issue Hillis and Bull (1993) present simulations on the behavior of the bootstrap method of assigning $\underline{P}$-values to the presence of groups in phylogenies, as advocated by one of us (Felsenstein, 1985). They find that the $\underline{P}$ values are less extreme, sometimes considerably so, than the probabilities that the groups are correct. Among groups that have a value of $\underline{P} = 0.85$, for example, over 95% of them may be correct groups. Thus the value of $\underline{P}$ is conservative in their simulations, sometimes extremely so. This raises the question of whether there is some flaw in the logic of the bootstrap.

We do not at all dispute the results of Hillis and Bull's simulation, which is a very thorough and helpful exercise. We want instead to comment on where the phenomena that they find come from. We will argue that they are not a result of using the bootstrap, but are a result of summarizing the evidence for a group by using a $\underline{P}$ value. We will show that their phenomena are rather precisely duplicated in a much simpler model of estimating where the mean of a normal distribution is, a model having no bootstrapping. As in empirical studies, we can often get a clearer picture by considering a simple example. Finally, we will show that there is another straightforward meaning of the $\underline{P}$ value that is not invalidated by Hillis and Bull's criticisms, and which can be taken as the "real" meaning of the bootstrap $\underline{P}$ value.

Recently, Zharkikh and Li (1992) have provided a careful and detailed theoretical treatment of the behavior of bootstrap $\underline{P}$ values in the case of four species. They have shown that the major phenomenon found by Hillis and Bull is expected theoretically. We will also show this in the simpler normal distribution model. Our analysis thus may provide some further understanding of the source of the phenomena found by Zharkikh and Li.

THEORETICAL BASIS

A Normal Distribution Example

Consider the problem of estimating the mean of a normal distribution, from a sample of $\underline{n}$ points drawn from it. We will simplify the problem by assuming that the variance of the distribution is known to be 1, with only the mean μ being unknown. As an analogy to the issue of tree topology, we will declare an interest in whether $\mu > 0$ or $\mu < 0$, with the point $\mu = 0$ being analogous to a multifurcation among tree topologies.

On observing a sample of $\underline{n}$ points with average $\underline{\bar{x}}$, our best estimate of μ is $\underline{\bar{x}}$, and our best estimate of the unknown true distribution of $\underline{x}$ is that it is a normal distribution centered at this value, with variance 1. It follows that our best estimate of the distribution of the mean $\underline{\bar{x}}$ is that it is drawn from a normal distribution with mean $\underline{\bar{x}}$ and variance $1/\underline{n}$. A straightforward way to assign a probability to the hypothesis that $\mu > 0$ is to take this latter normal distribution and see what fraction of its area lies in the region $\mu > 0$. This is the "obvious" way to assign a $\underline{P}$ value, and most of us would do this with little hesitation. Figure 1 shows this estimate, if the sample mean happened to be the same as the true mean.

Accuracy

The procedure seems to draw support from the observation that when we use the (unknown) true value of μ instead of $\underline{\bar{x}}$, $\underline{P}$ is the proportion of the time that $\underline{\bar{x}}$ would be in the correct "topology". If $\phi(\underline{x})$ is the standard normal density evaluated at $\underline{x}$, then noting that $\underline{\bar{x}}$ is normally distributed around a mean of μ with variance $1/\underline{n}$, we can calculate the probability that we get the right topology given a true mean of μ, which we call $\underline{P}(\mu)$, by

$$
\begin{aligned}
\underline{\text{Prob}}\ (\bar{\underline{x}} > 0) &= \tfrac{1}{\sqrt{1/\underline{n}}} \int_0^\infty \phi\left((\underline{x} - \mu) / \sqrt{1/\underline{n}}\right) dx \\
&= \int_{-\sqrt{\underline{n}}\mu}^{\infty} \phi(\underline{x}) \underline{dx} \qquad (1) \\
&= \underline{P}(\mu)
\end{aligned}
$$

as the value 0 is $\sqrt{\underline{n}}\mu$ standard deviations below the mean when the mean is μ and the variance is $1/\underline{n}$.

Thus the true $\underline{P}$ is the fraction of times, on repeated sampling, that the mean will come out above 0 so that we will correctly infer the "topology." This is "accuracy."

Repeatability

What of their "repeatability"? It too can be computed in this simple case. It is the average of the $\underline{P}$ values that we would compute from many different samples drawn from the true distribution. If a sample turns out to have a mean of $\bar{\underline{x}}$, the $\underline{P}$ value that we would calculate for it would be the probability of the observed mean being above 0 given that $\bar{\underline{x}}$ was the true mean:

$$
\underline{P}(\bar{\underline{x}}) = \int_{-\sqrt{\underline{n}\bar{\underline{x}}}}^{\infty} \phi(\underline{x}) \underline{dx} = \int_{-\infty}^{\sqrt{\underline{n}\bar{\underline{x}}}} \phi(\underline{x}) \underline{dx} = \Phi\left(\sqrt{\underline{n}\bar{\underline{x}}}\right) \qquad (2)
$$

where $\Phi(\underline{x})$ has its usual meaning of the area under the standard normal distribution below $\underline{x}$. Note that Equation 1 could also have been written as $\Phi(\sqrt{\underline{n}}\mu)$.

Note that although $\underline{\bar{x}}$ is distributed symmetrically around the true value μ, $\underline{P}(\underline{\bar{x}})$ is a nonlinear function of $\underline{\bar{x}}$. Thus although $\underline{\bar{x}}$ is an unbiased estimate of μ, $\underline{P(\bar{x})}$ is a biased estimate of $\underline{P}(\mu)$. This is exactly Hillis and Bull's observation. The $\underline{P}$ value computed given an observed mean of $\underline{\bar{x}}$ is the probability that $\underline{\bar{x}}' > 0$, where $\underline{\bar{x}'}$ is drawn from a distribution with expectation $\underline{\bar{x}}$ and variance $1/\underline{n}$, The mean $\underline{P}$ value averaged over all $\underline{\bar{x}}$ values is simply the probability that $\underline{\bar{x}'} > 0$. Since

$$\underline{\bar{x}}' - \mu = (\underline{\bar{x}}' - \underline{\bar{x}}) + (\underline{\bar{x}} - \mu) \tag{3}$$

and the two terms on the right-hand side of Equation 3 will each have expectation 0 and variance $1/\underline{n}$, and are independent, the difference $\underline{\bar{x}'} - \mu$ will have mean zero and variance twice as great as the variance of $\underline{\bar{x}}$. From this we can show that the probability of $\underline{\bar{x}'}$ being above 0 is

$$\bar{\underline{P}}(\underline{\bar{x}}) = \Phi\left(\frac{\sqrt{\underline{n}}\mu}{\sqrt{2}}\right). \tag{4}$$

This will always be closer to 0.5 than is the true $\underline{P}$. Figure 2 shows $\bar{P}$ plotted as a function of $\underline{P}$. When $\underline{P} = 0.95$ the value of $\bar{P}$ is only 0.86. Thus the average value of $\underline{P}$ that a user would get would be 0.86. This curve shows the relationship between "repeatability" (the horizontal axis) and our version of Hillis and Bull's "accuracy" (the vertical axis) for this simple normal distribution case. (Note that although we have used $\underline{n}$ in our derivations, both the true $\underline{P}$ and $\underline{\bar{P}}$ are functions of $\sqrt{\underline{n}}\mu$ and their dependence on each other is the same whatever the value of $\underline{n}$).

Zharkikh and Li (1992) have shown analytically for a case of four species with a molecular clock and simple models of DNA evolution that this bias is expected for bootstrap $\underline{P}$ values. Our normal distribution example thus reproduces the qualitative behavior of both their theoretical analysis and Hillis and Bull's simulation. It shows that Zharkikh and Li's analytical results for the bootstrap show a behavior which is not due to the presence of the bootstrap.

Variability of P values

Hillis and Bull also found that their $\underline{P}$ values for each case were highly variable. They will also be variable in the simple normal distribution. It can be shown that the density function of the estimated value of $\underline{P}$ will be given by

$$\exp\left(\sqrt{n}\mu\left(\Phi^{-1}(P) - \frac{1}{2}\sqrt{n}\mu\right)\right). \tag{5}$$

where $\Phi^{-1}(P)$ is the value of $\underline{x}$ which has $\Phi(x) = P$.

Figure 3 shows this density function, showing the distribution of $\underline{P}$ values for three different true values of $\sqrt{n}\mu$, 0, 1, and 2. Note the similiarity of these distributions to the histograms in Hillis and Bull's Figure 3. The $\underline{P}$ values in our simplified example are as variable as theirs. We return later to the curve for 0, which establishes a principle that illuminates the true meaning of the bootstrap $\underline{P}$ value.

Probability That a Group is True

Hillis and Bull also estimate the probability that a group declared to have a certain $\underline{P}$ value is actually a true group. In the simplified example we can compute the same quantity analytically. It depends on the true values of μ. In their case they chose true phylogenies by assigning branch lengths randomly. In our case we must have a true distribution of μ.

We will show the results for two cases. In the first the true μ is equally likely to be $+|\mu|$ and $-|\mu|$, where we specify the absolute value $|\mu|$. It is not hard to show that, given that we have computed a value $\underline{P}$ favoring the "topology" $\mu > 0$, the probability that the true value is in fact $+|\mu|$ is

$$1 \Big/ \left(1 + \underline{e}^{-2\sqrt{\underline{n}}|\mu|\Phi^{-1}(\underline{P})}\right). \tag{6}$$

Figure 4 shows the curves relating the $\underline{P}$ value to the probability of the "topology" being true for three values of $\sqrt{\underline{n}}|\mu|$, 0.1, 1.0, and 2.0. Note the similarity of these curves to the corresponding graphs in Hillis and Bull's figures 4, 5, and 6. In our Figure 4 the curves cross when the probability of a true result is 0.5, in theirs nearer 0.33, because they have multiple tree topologies while we have only two. As with their figures, the nearly flat curve is obtained when the alternative values of μ are so small that we have very little statistical power for estimating the true tree. In their case flat curves are obtained when the characters are distributed randomly among taxa, as would be the case when rates of change are so high that "noise" dominates "signal".

The other example that we can calculate is when μ is itself normally distributed with mean 0 and variance σ^2. The observed mean $\underline{\bar{x}}$ is the sum

$$\underline{\bar{x}} = (\underline{\bar{x}} - \mu) + \mu \tag{7}$$

with each of the two parts of the right-hand side of Equation 7 being normally distributed, with respective variances $1/\underline{n}$ and σ^2. The two pieces are also readily seen to be independent of each other. From this we can work out the distribution of μ conditional on $\underline{\bar{x}}$, which will be normal with mean

$$\underline{E}\left[\mu|\underline{\bar{x}}\right] = \left(\frac{\underline{n}\sigma^2}{1+\underline{n}\sigma^2}\right)\underline{\bar{x}} \tag{8}$$

and variance

$$\mathrm{Var}\left[\mu|\underline{\bar{x}}\right] = \frac{\sigma^2}{1+\underline{n}\sigma^2}. \tag{9}$$

This means that when we observe a mean $\underline{\bar{x}}$, we will calculate a $\underline{P}$ value of $\Phi(\sqrt{\underline{n}}\underline{\bar{x}})$, and the fraction of the distribution of μ conditional on $\underline{\bar{x}}$ which is above 0 will be

$$\text{Prob } \left[\mu > 0|\underline{\bar{x}}\right] = \Phi\left(\left(\frac{\underline{n}\sigma}{\sqrt{1+\underline{n}\sigma^2}}\right)\underline{\bar{x}}\right). \tag{10}$$

Figure 5 shows the curves relating Equation 10 to $\underline{P}$, plotted for three values of $n\sigma^2$, 0.1, 1, and 10. As with Figure 4, for true hypotheses that are not very different ($\underline{n}\sigma^2 = 0.1$) the probability of having the true "topology" is close to 0.5 for all values of $\underline{P}$, while when we have a great deal of information the probability of a true result is much higher than $\underline{P}$ when $\underline{P}$ is large ($\underline{n}\sigma^2 = 10$).

IMPLICATIONS

Both of the cases we have investigated thus end up showing the same pattern, that whether large $\underline{P}$ value over- or underestimates the probability that the group is correct depends on whether the data is highly informative or not. If the data have considerable power to resolve a portion of the phylogeny, the nominal $\underline{P}$ value will be an underestimate; in that case we will also expect to most frequently find large nominal $\underline{P}$ values.

Most importantly, we have reproduced the major phenomena of Hillis and Bull's simulations here in a simple case that does not involve any use of the bootstrap. We have been able to rely on the assumption that $\underline{\bar{x}}$ is distributed normally around a mean of μ with a known variance. In the much more complicated case of phylogeny estimation, we do not know how the estimated tree will be expected to vary around the unknown true tree, and it is the function of the bootstrap to give us an estimate of that distribution.

This it does quite well: the kinds and degree of variability that will be seen among bootstrap estimates of the tree will accurately reflect the kinds of variability we would see if we could sample new data sets from the same underlying true tree and estimate trees for each one, provided that we look at the details of the tree and not the $\underline{P}$ value . In the present normal distribution case, if we did not understand the distribution of $\underline{\bar{x}}$ and tried to use the bootstrap, it would consist of resampling $\underline{n}$ times with replacement from our $\underline{n}$ data points and computing the mean $\underline{\bar{x}}$ for each bootstrap sample. The estimated $\underline{P}$ value for the "topology" $\mu > 0$ would be the fraction of times we would get $\underline{\bar{x}} > 0$. This will vary between samples of $\underline{n}$ data points. When $\underline{n}$ is large the distribution of the $\underline{n}$ data points will be close to normal with mean $\underline{\bar{x}}$ and a variance which is drawn from a multiple of a χ^2-distribution with an expectation of $1/\underline{n}$. As we consider cases with larger and larger sample sizes, we can show that the bootstrap estimate of the distribution will approach the true distribution, as it should.

<u>**Confidence Intervals**</u>

Instead of computing $\underline{P}$ for the hypothesis that $\mu > 0$, we could have made a confidence interval on the value of μ. On finding a sample mean of $\underline{\bar{x}}$ the two-sided 95% confidence interval would be $(\underline{\bar{x}}-1.96/\sqrt{\underline{n}},\ \ \underline{\bar{x}}+1.96/\sqrt{n})$. Note that this differs from obtaining the $\underline{P}$ value by holding $\underline{P}$ constant and varying the interval, whereas what we have been talking about is holding the interval constant and computing different values of $\underline{P}$ with each sample. We shall see below that when we make a confidence interval, there turns out to be no bias at all. The probability that our confidence interval will include the true value of μ is 0.95, exactly as it should be. If the bootstrap were used instead of knowing the variance, the confidence interval would be a bit too large on average but as the amount of data were increased it would rapidly become correct.

However this does not end up guaranteeing that the $\underline{P}$ value for a fixed interval will be unbiased, as we have seen. The correctness of the confidence interval computed from the bootstrap is not questioned by the Hillis and Bull results. If we could have some method of summarizing the results that displayed the confidence interval, rather than simply assigning a $\underline{P}$ value, this would avoid most of the problems and provide a method that would provide systematists and molecular evolutionists with an accurate assessment of the statistical uncertainty in their phylogenies.

Bootstrap Confidence Sets

We can test some of the above statements about the accuracy of the bootstrap confidence interval by computer simulation. A good confidence region is small and has an unbiased coverage probability, in the sense that the true parameter value will be included in the confidence region a fraction of times which is expected to be equal to the prespecified $\underline{P}$ value (the nominal level).

In the inference of phylogenies, the notion of a confidence region must be generalized. It is a set of tree topologies or trees which we will refer to as a confidence set. Having different tree topologies is similar to having several different statistical models. It is not obvious what criterion to use to delimit confidence sets. In the simulation studies here, we have used the likelihood ratio between the true model and the other possible models. We have used bootstrap sampling to characterize the distribution of this likelihood ratio and determined confidence sets based on that.

Numerical Study of N(μ, 1) vs N(0, σ^2)

We look at the coverage probabilities of the above confidence sets in a simple example. Two models of normal distributions, N(μ, 1) and N(0, σ^2), are compared here. The first model has the mean μ unknown, but the variance known, and the second model has the variance σ^2 unknown but the mean known. This pair of models thus looks something like a pair of tree topologies: each model has its own parameters just as each topology has its own branch lengths. A thousand sets of independent and identically distributed samples of size 100 from one of these distributions were generated. For each sample, confidence sets with nominal coverage probabilities 0.50, 0.75, 0.80, 0.85, 0.90 and 0.95 were constructed using 1,000 bootstrap replicates from each data set. In each the likelihood ratio between the true hypothesis and the best estimate from that bootstrap sample was computed. The distribution of these likelihood ratios was then used to determine how closely, as measured by the likelihood ratio, the estimated values were likely to be to the true values. We could determine a threshold value of the likelihood ratio that could be used to exclude hypotheses having a lower likelihood from the confidence set. We could then see for each of our 1,000 original samples whether the likelihood ratio between the true values and the sample values was greater than this threshold. We observe the proportion among the 1,000 confidence sets that cover the true parameter value (for the confidence sets on the parameters).

Five values of $\mu = -0.2, -0.1, 0.0, 0.1, 0.2$ were studied in the first model, and also five values of $\sigma^2 = 0.6, 0.8, 1.0, 1.25, 1.67$ were studied in the second model. As is expected, the repeatability is 0.5 when the true model is N(0, 1) (the boundary between the "topologies"). Calculated coverage probabilities are plotted against the nominal ones in Figure 6. The different points at each $\underline{P}$ value are for the 10 different true parameter values tried. We see that the confidence sets of both types have fairly unbiased coverage probabilities.

In the case of phylogenies it is not obvious how best to construct the confidence sets, especially when many tree topologies are possible. Nor is it easy to see how best to construct them so that they are biologically meaningful to the user. Other ways of summarizing the variation seen in the bootstrap estimates of the phylogenies have been suggested. Sanderson (1989) has suggested reporting the size of monophyletic groups in which the group of interest is contained in 95% of the bootstrap estimates of the phylogeny. This suggestion makes a confidence interval rather than reporting a $\underline{P}$ value. It thus would not be subject to the biases reported by Hillis and Bull; its main disadvantage is that users may be less interested in its report than they would be in an assessment of the amount of support for a predefined group. Note that when a particular monophyletic group is supported 95% of the time, it will be reported identically by both Sanderson's method and the original method.

A Modest Proposal

Note that although Figures 4 and 5 seem to say that in some cases a $\underline{P} = 0.95$ might be a gross overestimate, those will be the cases where the support for that topology will on average be weak. Thus in those cases we will rarely find $\underline{P}$ values that high; if we do, they will grossly overestimate the probability of the group being correct, but mostly the $\underline{P}$ values will come out lower so that we will not often be faced by the problem. In fact the density shown in Figure 3 is the distribution of $\underline{P}$ values: note that when $\mu = 0$ the probability of obtaining a value of $\underline{P} > 0.95$ is exactly 0.05. If $\underline{P} > 0.95$ is used as the criterion, when $\mu > 0$ the probability of concluding falsely that there is significant evidence that $\mu < 0$ is easily shown to also always be less than 0.05.

If we compute the probability that there is evidence favoring a group giving a nominal probability of $\underline{P}$, the probability that this much evidence will favor the group when it is not in fact on the true tree is less than or equal to $1 - \underline{P}$. This, we think, is the right way to think about the $\underline{P}$ value: $1 - \underline{P}$ tells us the probability of Type I error, of falsely favoring the group under the null hypothesis that it is not there. Of course the probability $1 - \underline{P}$ is correct only in the worst case, when the true tree has a multifurcation so that the group is almost correct. Otherwise it is conservative.

We suggest that when systematists see the group of interest occur a fraction $\underline{P}$ of the time among bootstrap samples, that they should regard $1-\underline{P}$ as a conservative assessment of the probability of getting that much evidence favoring the group if it is not present. A small value of $1-\underline{P}$ will then, like the tail probability in a hypothesis test, make more likely the rejection of the null hypothesis that the group is absent. It will be a conservative procedure, but one with a straightforward interpretation. For example, when the group is seen to have $\underline{P} = 0.85$, that much evidence favoring the group would be expected less than 15% of the time if the group were not on the true tree.

Hillis and Bull suggest that one possible reaction is suspending all belief in the $\underline{P}$ value except as a measure of relative confidence for different groups in the same tree. We think that this would be an overreaction, and that biologists will continue to use the $\underline{P}$ values (until a better way of summarizing the results is developed). Thanks to Hillis and Bull's helpful work they may do so with some clarification of the meaning of the $\underline{P}$ value.

ACKNOWLEDGEMENTS

This work was supported by NIH grants 2R01GM41716 and 2R55GM41716, and by NSF grants BSR-8918333 and DEB-92070558. We thank Sean Lamont for assistance with the simulations and Mary Kuhner, Jon Yamato, Peter Guttorp, Elinor Ytterstad and Jeffrey Thorne for helpful discussions.

REFERENCES

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a measure of assessing confidence in phylogenetic analysis. Syst. Biol., in press.

Sanderson, M. J. 1989. Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5:113-129.

Zharkikh, A., and W.-H. Li. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9:1119-1147.

Figure Captions

Figure 1. Diagram to illustrate the normal distribution model, with the distribution (dashed lines), the two regions ("topologies"), the distribution of sample means, and the fraction of the estimated distribution that would fall above 0 (and hence be used to obtain $\underline{P}$) if the sample mean happened to be at the true mean.

Figure 2. Average value of $\underline{P}$ that would be calculated for true distributions that had different true values of $\underline{P}$, the true value being the probability that a new sample from that distribution would be above 0.

Figure 3. Distributions of the value of $\underline{P}$ obtained in different samples, for various true values of the mean μ of the distribution. These distributions show the "precision" of the estimation of $\underline{P}$.

Figure 4. The probability that the "topology" obtained is correct as a function of the calculated $\underline{P}$ value for three different values of $|\mu|$, where the truth is equally likely to be $+\mu$ and $-\mu$.

Figure 5. The probability that the "topology" obtained is correct as a function of the calculated $\underline{P}$ value for three different values of $\underline{n}\sigma^2$, where the true value of μ is distributed normally around 0 with a variance of σ^2.

Figure 6. Calculated coverage probabilities obtained in a computer simulation when the likelihood ratio between models is used to select an interval estimate from models which have either different means or different variances when data are simulated from some of these models, and the bootstrap used to decide how large a likelihood ratio to allow. The horizontal axis is the nominal $\underline{P}$ value, the vertical axis the fraction of times that it is found that the interval includes the true model. Further details in text.