

Lecture Outline: Assessing Uncertainty with the Bootstrap

1. The Bootstrap

- The bootstrap was introduced to the world by Brad Efron, chair of the Department of Statistics at Stanford University, in 1979.
- The bootstrap is one of the most widely used new method in statistics that was invented within the past 50 years.
- In a special issue of *Statistical Science* that celebrates the 25th anniversary of the bootstrap, Brad Efron uses its application to phylogenetics as one of a small number of examples to illustrate its use and importance.

The general principle:

- (a) We have a sample x_1, \dots, x_n drawn from a distribution F from which we wish to estimate a parameter θ using a statistic $\hat{\theta} = T(x_1, \dots, x_n)$.
(We might think of θ as being the median of the distribution, for example, and $\hat{\theta} = T(x_1, \dots, x_n)$ as the sample median.)
- (b) If we wanted to compute the standard error of the estimate, we would ideally compute the standard deviation of $T(X_1, \dots, X_n)$ where $X_i \sim \text{iid } F$.
- (c) We could estimate this to any desired degree of accuracy by generating a large enough number (say B) of random samples X_1, \dots, X_n , computing $T_i = T(X_1, \dots, X_n)$ for the i th such sample, and then computing the standard deviation of these estimates.

$$\sqrt{\frac{\sum_{i=1}^B (T_i - \theta)^2}{B}}$$

- (d) Unfortunately, we cannot take multiple samples from F .
- (e) However, our original sample x_1, \dots, x_n is an estimate of the distribution F .
- (f) Instead of taking samples from F , we could sample from the estimated distribution \hat{F} by sampling from our original sample *with replacement*.
- (g) We sample n values x_1^*, \dots, x_n^* with replacement from x_1, \dots, x_n . It is very likely that some of the original x values will be sampled multiple times and others will not be sampled at all.
- (h) For each sample, compute the estimate of θ using the original statistic. The i th estimate is $T_i^* = T(x_1^*, \dots, x_n^*)$.
- (i) Repeat this B times and compute the standard deviation of the bootstrap estimates around the estimate from the original sample.

$$\sqrt{\frac{\sum_{i=1}^B (T_i^* - \hat{\theta})^2}{B}}$$

- (j) If the sampling distribution of the bootstrap sample estimates around the estimate $\hat{\theta}$ is similar to the sampling distribution of the estimates $\hat{\theta}$ around the true value θ , then the bootstrap standard error will be a good estimate of the real standard error.

2. Applying the Bootstrap in Phylogenetics

- (a) The bootstrap was introduced in phylogenetics by Joe Felsenstein in 1985.
- (b) The bootstrap may be applied to any tree reconstruction method that produces a single estimated tree from a sample of data (such as maximum likelihood, neighbor-joining, UPGMA, and maximum parsimony).
- (c) Assume that we have n sites in an alignment of DNA sequences.
- (d) Treat the sites as an independent sample.
- (e) Estimate the tree using your method of choice from the original alignment.
- (f) Create B bootstrap samples, each of which consists of n sites selected uniformly at random with replacement from the original alignment of n sites.

- (g) For each bootstrap data set, use the original estimation procedure to estimate the phylogeny, creating B bootstrap trees.
- (h) For each clade in the original estimate, report the proportion of bootstrap trees that contain the clade.
- (i) Alternatively, use a consensus method to summarize the sample of bootstrap trees.

3. Consensus Trees

- (a) A *strict consensus tree* shows only those clades that appear in every sampled tree.
- (b) A *majority rule consensus tree* shows all clades that appear in more than half the sample of trees. (Notice that two clades that each appear in more than half the sampled trees must appear in at least one tree together, implying that they are compatible with one another.)
- (c) A *priority consensus tree* adds clades to the majority rule consensus tree in order of decreasing frequency in the sample provided that these clades do not conflict with a clade with higher frequency.

4. Dynamic Exploration of Tree Samples

Show off Mark Derthick's **Summary Tree Explorer**.

Software is free and available at <http://cityscape.inf.cs.cmu.edu/phylogeny/>.

5. Interpretation of Bootstrap Proportions

What does a bootstrap proportion mean? Let me count the ways.

- (a) *Confidence* that the clade is in the true tree.
- (b) Bayesian posterior probability that the clade is in the true tree.
- (c) One minus p-value for a formal hypothesis test that the clade is in the true tree.
- (d) Rough measure of method robustness.
- (e) Measure of repeatability of the inferences for the method at hand.
- (f) Others?