Phylogenetics Z620 Homework 8

In this homework we are going to explore the bias in the maximum likelihood estimator of the rate of an exponential distribution.

Facts:

(1) The density of the exponential distribution with rate λ is

$$f(x) = \lambda e^{-\lambda x}$$

(2) Given data, x_1, x_2, …. from an exponential distribution the maximum likelihood estimator for the rate, lambda (λ), is 1 / average(x).

(3) If there is a large enough sample size, this maximum likelihood estimator is approximately normally distributed. If there is only a small sample, then this maximum likelihood estimator is skewed.

Instructions:

Download the Excel File with the macro from
http://mypage.iu.edu/~ehouswor/Fall2005/BioZ620/Bootstrap.xls

Under Tools -> Macros -> Macro, run the only macro in the file. This will cause a dialog box to pop up.

Always choose to perform at least 1000 bootstraps. The effect of increasing the number of bootstraps is simply to get more accurate estimates for the ends of the confidence intervals. Choosing to use many confidence intervals, similarly, gives you accurate estimates for the actual confidence obtained. The interesting parameter to vary is the sample size.

(1) First let the sample be of size 1. The code then does the following: For each confidence interval, it chooses a sample of size 1 from an exponential distribution with rate λ=1. It estimates λ as 1/x. For every one of the bootstraps, it simulates a sample of size 1 from an exponential distribution with rate (1/x) and re-estimates λ. It finds the 95% limits on the re-estimates of lambda and reports them (using either the percentile method or Rice's method.)

Using 1000 bootstraps and 1000 intervals, what is the proportion of times the CI actually covers 1? It should be 95% but it is actually much lower.

Look at the distribution of estimates for lambda reported in column 3 of Sheet1. If you have "DataAnalysis" under Tools in Excel, use this to histogram the values of lambda. The values are in column C. If you have 1000 intervals, then the values go from C2:C1001. You can store the output in either a separate sheet or in range K1:L100 (excess range is ok here). You can also form a column of bin end-points and include that column in "Bin" which is useful for creating smaller bins for the skewed distributions than Excel creates by default. Or you can export the data into another program and create the histograms there.

(2) Now let the sample size slowly increase through 2, 3, 4, 5. Look at the distribution for the estimates of the rates λ as above. Look at the achieved level of coverage as reported in the Excel spreadsheet. Which works better, Rice or the percentile bootstrapping method? When (for what sample size) is nominal 95% coverage (approximately) achieved?