Bootstrap and Resampling Methods, spring 2023

Homework exercise 3

Due date: 28 May 2023 before class

1. **Problem 16.15 from the book: A different approach to testing the mean.** Suppose we have a sample $z_1, ..., z_n$ and we want to estimate the underlying distribution F restricted to have mean μ. One approach, used in Sec. 16.4, is to use the empirical distribution of the translated data values $z_i - \bar{z} + \mu$. A different approach is to leave the data values fixed, and instead change the probability p_i on x_i for each i to be different than $1/n$. Let $\mathbf{p} = (p_1, ..., p_n)$ and let F_p be the "empirical" distribution putting probability p_i on x_i. Then we want to choose p such that the mean of $F_p = \sum_i p_i x_i = \mu$ and F_p is as close as possible to the empirical distribution $\hat{F}$. A convenient measure of distance is the Kullback-Leibler distance

$$d_{F_p}(F_p, \hat{F}) = \sum_{i=1}^{n} p_i \log\left(\frac{1}{np_i}\right).$$

   (a) Using Lagrange multipliers or any other way, show that the probabilities that minimize this distance subject to $\sum p_i x_i = \mu$ and $\sum p_i = 1$ are given by

$$p_i = \frac{\exp(tx_i)}{\sum_i \exp(tx_i)}$$

   when t is chosen such that $\sum p_i x_i = \mu$. What can you say about t when $\mu > \mathbf{E}(\hat{F})$, and when $\mu < \mathbf{E}(\hat{F})$? Interpret the resulting F_p.

   (b) Use this approach to carry out a test of $\mu = 129$ on the mouse treatment data (`mouse.t` object in the package `bootstrap`). Compare the results to those obtained with the "standard" approach in Section 16.4.

2. **Out of bag bagging and random forest**

   (a) Both algorithms include bootstrap sampling of observations in each of the iterations (for example, when building each tree). Show that the probability of each observation to be selected into the sample is approximately $1 - 1/e$.

   (b) (* Extra credit) Can you calculate the approximate probability of each observation to be selected exactly once in the sample?

   (c) The observations not selected in each iteration are called "out of bag" and can be used as a "test set" for this tree. Suppose I want to use this fact to evaluate the performance of

a random forest or bagging algorithm using 1000 trees on holdout data, without having a separate test set. Explain clearly how I can use the out of bag data to do this, and how many trees I would actually need to build (relate this to the probability in part (a).

3. **Questions on phylogenetic inference presentation.**

   (a) **The Dirichlet prior argument.**
   Consider a simplified version of the situation in slides 41-44, where instead of multinomial we have a binomial with $\pi = (0.3, 0.7)$ and $n = 50$. The two-value version of Dirichlet is the Beta distribution.

      i. Draw three samples of size 100:
         - From Binomial(50,0.3).
         - From the appropriate Beta posterior when $\alpha_1 = \alpha_2 = 0$.
         - From the Beta posterior when $\alpha_1 = \alpha_2 = 10$.

      Draw the empirical cumulative distribution for the three samples and comment on their similarities relative to Efron's claims.

      ii. Consider Efron's model of the process as:

$$\pi \to D \to \text{Tree} \to \psi,$$

      where D is the distance distribution. What part of this model do we need to generate the sequences that are our observed data (hint: consider the meaning of π carefully)? Does this model make sense as describing the way the "world works" and the data are truly generated?

      iii. (* Extra credit) Explain how the problem with this model implies that the prior which has $\alpha_1 = ... = \alpha_K = 0$ does not make sense for this application

   (b) **The final algorithm.**
   Consider the algorithm Efron proposes on slides 60-65.

      i. Explain the goal of the second bullet on slide 61 and relate it to the methodology for hypothesis testing with bootstrap. In particular, explain the statement in the third bullet. Propose another way to accomplish this goal.

      ii. By analogy to the normal example given before, explain the meaning of the ratio between the second and third columns in the table on slide 62 — how does it relate to the shapes in Fig. 5 on slide 54? What does it mean for Felsenstein's p-value — is it too small or too big? Explain.