Lecture 28: Random Numbers: Bootstrapping

BT 3051 – Data Structures and Algorithms for Biology

Karthik Raman
Department of Biotechnology
Indian Institute of Technology Madras

Why Bootstrapping?

Traditional Statistical Inference

- From any sample, we can figure only *one* of each statistic Ω (mean, variance, max, ...)
- But, what is the distribution of Ω (e.g., shape, variance, skewness, etc.)
- Using mathematical proofs that 99.9% of users do not understand, traditional inference infers the distribution of Ω given that *certain assumptions* are met!
- *p*-values and confidence intervals are suspect when assumptions are violated!

Courtesy: Excellent course on "Learn R" from Colorado

Bootstrapping

Bootstrap seeks to find the distribution of any possible Ω

- without assumptions about the population distribution
- without deriving the sampling distribution explicitly
- *only from knowledge of the sample itself* — i.e. the sample pulls itself up by its bootstraps

Bootstrapping

- Take a sample S_1^* of size n from your original sample S *with replacement*
  - If we sampled without replacement, we would just recreate S!
- Calculate your statistic of interest, Ω_1^*
- Repeat the above two steps several (n) times, each time saving $\Omega_1^*, \Omega_2^*, \ldots, \Omega_n^*$
- The distribution of the n Ω^*'s is a good estimate of the true distribution of Ω

Assumptions of Bootstrapping

- Bootstrapping is *nearly* assumption free!
- The key assumption is that the distribution of your sample, S, is a close approximation to the population distribution
  - tends to be true as *n* becomes large
- But if *n* is small, this is precisely when violations of assumptions for traditional statistics matters most , but ...
  - inference will always be suspect for small samples!
  - no statistical gymnastics can make that go away

Bootstrap obviates the need for assumptions that had to be made before modern computing — and it allows us to estimate *unknowable* sampling distributions

Applications of Bootstrapping

- Phylogenetic Trees (from sequence data)
- Gene interactions (from microarray data)
- ...